SOLID-STATE ELECTRONICS

AN INTERNATIONAL JOURNAL

VOLUME 2, 1961

Editor-in-Chief
W. CRAWFORD DUNLAP
Waltham, Mass.

Regional Editors

U.K.
L. PINCHERLE
London

EUROPE
H. WELKER
Erlangen, Germany

JAPAN and FAR EAST
G. M. HATOYAMA
Tokyo

PERGAMON PRESS
OXFORD · LONDON · NEW YORK · PARIS

SOLID-STATE ELECTRONICS

AN INTERNATIONAL JOURNAL

VOLUME 2, 1961

Editor-in-Chief
W. CRAWFORD DUNLAP
Waltham, Mass.

Regional Editors

U.K.
L. PINCHERLE
London

EUROPE
H. WELKER
Erlangen, Germany

JAPAN and FAR EAST
G. M. HATOYAMA
Tokyo

PERGAMON PRESS
OXFORD · LONDON · NEW YORK · PARIS

LIST OF CONTENTS

NUMBER 1

L. M. TERMAN: Spectral response of solar-cell structures 1

F. E. ROBERTS: Impurity concentrations in the regrown region of In–Ge alloyed junctions 8

B. JANSEN: A rapid and accurate method for measuring the thickness of diffused layers in silicon and germanium 14

G. W. A. DUMMER: Recent advances in microminiaturization, reliable components and cooling techniques 18

WILLIAM SHOCKLEY: Problems related to *p–n* junctions in silicon 35

Notes:

T. KAN: Field-effect pinch-off of surface leakage in high-voltage diodes . . 68

S. DENDA: Micromasking for chemical etching of semiconductors 69

J. L. PARMEE and J. A. ROBERTSON: Programming the zone-refining equation for a digital computer 70

D. R. FRANKL: Some effects of material parameters on the design of surface space-charge varactors 71

Book Review:

Silicon Carbide—A High Temperature Semiconductor, Edited by J. R. O'CONNOR and J. SMILTENS (Reviewed by L. PINCHERLE) 77

Erratum:

L. A. D'ASARO: Diffusion and oxide masking in silicon by the box method . . 78

NUMBERS 2/3

J. LINDMAYER and C. WRIGLEY: The high-injection-level operation of drift transistors . 79

W. ALBERS: Diffusion of arsenic in germanium from the vapour phase . . . 85

W. G. SPITZER, J. H. WERNICK and R. WOLFE: Electrical and optical properties of $CdSnAs_2$ 96

ROBERT H. REDIKER and ALAN L. MCWHORTER: Compound cryosars for low-temperature computer memories 100

A. TRAINOR and B. E. BARTLETT: A possible mechanism of crystal growth from the melt and its application to the problem of anomalous segregation at crystal facets . 106

R. N. HALL: An analysis of the performance of thermoelectric devices made from long-lifetime semiconductors 115

EILEEN TANNENBAUM: Detailed analysis of thin phosphorus-diffused layers in *p*-type silicon 123

HANS P. KLEINKNECHT: Indium arsenide tunnel diodes 133

A. K. JONSCHER: Notes on the theory of four-layer semiconductor switches . . 143

J. G. VAN SANTEN and G. DIEMER: Photorectifier based on a combination of a photoconductor and an electret 149

NUMBERS 2/3—*continued*

GWYNN MCCONNELL and ROBERT SEHR: Measurement of the electrical resistance of metal–thermoelectric semiconductor contacts 157

G. T. WRIGHT: Mechanisms of space-charge-limited current in solids . . . 165

Notes:

W. RINDNER and J. M. LAVINE: Contact-etching of semiconductors . . . 190

B. L. H. WILSON: Control of lifetime and diffusion depth in semiconductor devices . 191

Book Review:

Proceedings of the Fourth Conference on Carbon, Edited by S. MROZOWSKI (Reviewed by R. MANSFIELD) 19-

NUMBER 4

S. MAYBURG: Direct recombination in GaAs and some consequences in transistor design 195

S. L. MATLOW and E. L. RALPH: A low-resistance ohmic contact for silicon semiconductor devices 202

T. P. BRODY and R. H. BOYER: The evaluation of "Esaki integrals" and an approximate expression for the tunnel-diode characteristic 209

SHOJI TAUCHI: Molten–diffused silicon transistors 216

T. S. MOSS: The potentialities of silicon and gallium arsenide solar batteries . . 222

ALBRECHT G. FISCHER: Injection electroluminescence 232

J. LINDMAYER and C. WRIGLEY: Alpha cutoff frequency of junction transistors . . 247

Notes:

C. T. THOMPSON and J. F. GERBER: Superconducting vanadium thin films . 259

KURT LEHOVEC: Distribution of thermal-conversion centers in germanium with indiffused impurities 260

Papers to be published in future issues. 262

AUTHOR INDEX

(B.R. indicates a Book Review)

ALBERS W., 85
BARTLETT B. E., 106
BOYER R. H., 209
BRODY T. P., 209
D'ASARO L. A., 78
DENDA S., 69
DIEMER G., 149
DUMMER G. W., A. 18
FISCHER A. G., 232
FRANKL D. R., 71
GERBER J. F., 259
HALL R. N., 115
JANSEN B., 14
JONSCHER A. K., 143
KAN T., 68
KLEINKNECHT H. P., 133
LAVINE J. M., 190
LEHOVEC K., 260
LINDMAYER J., 79, 247
MCCONNELL G., 157
MCWHORTER A. L., 100
MANSFIELD R., 194 (B.R.)
MATLOW S. L., 202
MAYBURG S., 195
MOSS T. S., 222
PARMEE J. L., 70
PINCHERLE L., 77 (B.R.)
RALPH E. L., 202
REDIKER R. H., 100
RINDNER W., 190
ROBERTS F. E., 8
ROBERTSON J. A., 70
SANTEN J. G. VAN, 149
SEHR R., 157
SHOCKLEY W., 35
SPITZER W. G., 96
TANNENBAUM E., 123
TAUCHI S., 216
TERMAN L. M., 1
THOMPSON C. T., 259
TRAINOR A., 106
WERNICK J. H., 96
WILSON B. L. H., 191
WOLFE R., 96
WRIGHT G. T., 165
WRIGLEY C., 79, 247

Solid-State Electronics Pergamon Press 1961. Vol. 2, No. 1, pp. 1–7. Printed in Great Britain

SPECTRAL RESPONSE OF SOLAR-CELL STRUCTURES*

L. M. TERMAN

Stanford Electronics Laboratories, Stanford University, California

(*Received* 22 *April* 1960; *in revised form* 24 *June* 1960)

Abstract—The purpose of this study was to measure the spectral response of silicon solar-cell structures, and to observe how the response varied with the depth of the *p–n* junction. Spectral response was defined as the relative short-circuit current as a function of the wavelength of incident light for equal energy incident upon the cell at all wavelengths. Cells were made with the junction depths varying from 0·6 to 5·0 μ, and having both smooth (etched) surfaces and rough (lapped) surfaces. Response curves from these cells are presented. These curves indicate that in order to increase the relative short-wavelength response ($\lambda < 0{\cdot}75\ \mu$) the junction should be made closer to the surface, while in order to increase the relative long-wavelength response ($\lambda > 0{\cdot}75\ \mu$) the junction must be made comparatively far below the surface. The effect of having a lapped surface on the cell is to reduce the lifetime near the surface, thus reducing the response to short wavelengths of incident light. A simple theoretical model is presented which appears adequately to describe the mechanism involved in determining the shape of the response curves.

Résumé—L'objet de cette étude est de mesurer la réponse spectrale des structures de cellules solaires au silicium et d'observer comment cette réponse changeait en fonction de la profondeur de la jonction *p–n*. La réponse spectrale est définie comme le courant relatif à court-circuit en fonction de la longueur d'onde de la lumière incidente (la quantité d'énergie incidente sur la cellule à toutes les longueurs d'onde étant égale). Les cellules ont été fabriquées en ayant des profondeurs de jonction variant de 0,6 à 5,0 μ, et en ayant des surfaces lisses (gravées) et des surfaces rugueuses (jointes). Les courbes de réponses de ces cellules sont ensuite présentées. Les courbes de ces réponses démontrent que pour augmenter la réponse relative de la longueur d'ondes courtes ($\lambda < 0{,}75\ \mu$) la jonction doit être faite plus près de la surface tandis que pour augmenter la réponse relative de la longueur d'ondes longues ($\lambda > 0{,}75\ \mu$), la jonction doit être faite loin de la surface. L'effet d'avoir sur la cellule une surface jointe est de réduire la durée de vie près de cette surface réduisant la réponse aux longueurs d'ondes courtes de lumière incidente. Un simple modèle théorique, qui semble indiquer d'une manière suffisante le mécanisme engagé pour déterminer la forme de ces courbes est ensuite présenté.

Zusammenfassung—Der Zweck dieser Untersuchung war die Messung der spektralen Empfindlichkeit von Silizium-Sonnenzellen, und die Beobachtung der Abhängigkeit der Empfindlichkeit von der Tiefe des *p–n*-Übergangs. Die Empfindlichkeit wurde mittels der Abhängigkeit des relativen Kurzschluss-Stroms von der Wellenlänge des einfallenden Lichtes gemessen, wenn bei allen Wellenlängen derselbe Energiebetrag auf die Zelle auftrifft. Die Tiefe des Übergangs schwankte zwischen 0,6 und 5 μ, und die Oberflächen waren entweder glatt (geätzt) oder rauh (geläppt). Die erhaltenen Kurven zeigen, dass zur Erhöhung der Kurzwellenempfindlichkeit ($\lambda < 0{,}75$) der Übergang näher an der Oberfläche liegen muss, während zur Erhöhung der Langwellenempfindlichkeit ($\lambda > 0{,}75$) der Übergang verhältnismässig tief unter der Oberfläche liegen muss. Der Effekt der geläppten Oberfläche der Zelle besteht darin, die Lebenszeit nahe der Oberfläche zu verkürzen, wodurch die Kurzwellenempfindlichkeit vermindert wird. Es folgt die Besprechung eines theoretischen Modells, das den Mechanismus zur Bestimmung der Gestalt der Empfindlichkeitskurven in angemessener Weise wiederzugeben scheint.

I. INTRODUCTION

A SOLAR cell is a semiconductor device which uses a *p–n* junction to convert the energy of the sun's rays directly into electrical energy.[1,2] The active surface of cells currently in use consists of a thin layer of *p*-type silicon on top of a body of *n*-material. The purpose of this study was to

* This work was supported through ONR Contract Nonr 225(31) at Stanford Electronics Laboratories, Stanford University.

measure the spectral response (here defined as the short-circuit current as a function of the wavelength of incident light) of solar cells, and to observe how it changed with the depth of the p–n junction. Measurements were made on cells with both smooth (etched) and rough (lapped) surfaces.

II. CONSTRUCTION OF CELLS

The steps employed in making the cells tested were as follows: a silicon slice (anywhere from 15 to 25 mils thick) of about 0·5 Ω-cm resistivity was lapped by hand using 600 carborundum grit to eliminate any surface irregularities. Cells which were to have lapped surfaces were cleaned in hydrofluoric acid. Cells which were to have etched surfaces were etched to a mirror finish by an immersion with agitation in a mixture of 5 parts nitric to 1 part hydrofluoric acid for about 2 min. The slices were then diffused in vaporous boron oxide at 1100°C for the required length of time. Upon removal from the furnace any oxide was removed with an HF dip. The p-material on one side was lapped away. On the other side an island of p-material was masked with wax, and the rest of the p-material was etched off. This insured a good p–n junction where the junction met the surface of the cell.

The contact to the n-material was made by evaporating gold onto the slice and then alloying at 500°C for 5 min. The contact to the p-material was made by evaporating an aluminum dot onto the slice and then alloying for 5 min at 700°C. The aluminum alloys into the p-material to a depth in excess of the several-micron junction depth but, since the aluminum is itself an acceptor-type impurity in silicon, the p–n junction is not destroyed but rather is simply moved deeper into the silicon slice in the vicinity of the dot. The finished cell is shown in Fig. 1(a) and (b).

The cells made by the above procedure differ from commerical cells in several important aspects. A commercial cell has a black skin on the active surface which may improve performance by reducing reflection. Further, such cells have a rough surface resulting either from lapping or sandblasting. Lapping (or sandblasting) causes disturbances in the surface to perhaps a 10-μ depth. Since this is several times a typical junction depth of 2 μ, it can be seen that there must be considerable doubt as to what surface conditions do exist in a commercial cell.

To avoid confusion as to surface conditions, the bulk of the spectral measurements were made on units having surfaces etched to a mirror finish to reduce surface damage to a minimum.* For comparison, some measurements were made on cells with lapped surfaces, but in all cases the surfaces were kept as clean as possible.

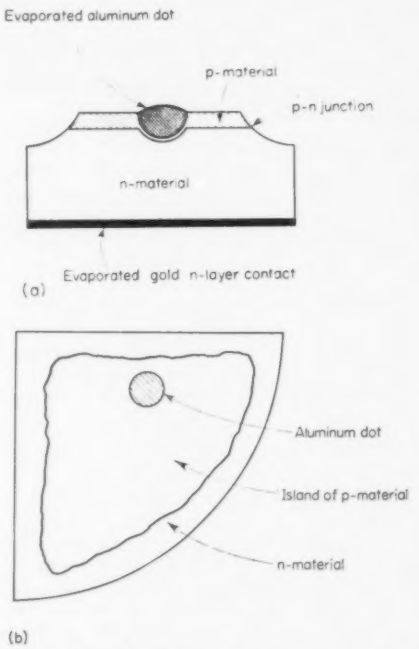


FIG. 1. Solar-cell configuration. (a) Cross-sectional view. (b) Top view.

The electrical characteristics of the cells upon which the spectral measurements were made are given in Table 1, and similar characteristics for a commercial cell are included for comparison. The open-circuit voltage and short-circuit current were measured in bright sunlight, and were normalized to a light intensity of 1000 W/m^2.

The junction-depth measurements on etched surfaces were made as follows: a shallow groove

* The mirror finish will increase losses due to reflection. It is felt that this will account for the greater short-circuit current density of the commercial cell when compared to the best etched cells, as seen in Table 1.

Table 1

	Diffusion time at 1100°C (min)	Junction depth (μ)	Open-circuit voltage (V)	Short-circuit current (mA/cm^2)
Etched surfaces	6	0·6	0·52	15·5
	12	1·0	0·52	13·9
	18	1·35	0·51	11·3
	24	1·7	0·48	12·9
	36	2·0	0·51	11·3
	72	3·2	0·44	7·5
	144	4·0	0·48	5·9
	288	5·0	0·50	3·75
Lapped surfaces	6	0·6*	0·32	10·5
	12	1·0*	0·32	8·6
	18	1·35*	0·30	10·0
	24	1·7*	0·40	9·4
	36	2·0*	0·33	8·6
	72	3·2*	0·20	4·6
Typical commercial cell	18		0·55	28·0

* Estimated junction depth from diffusion time.

of varying depth was lapped into the surface of the cell by a rotating bar held at a slight angle to the surface of the cell. Aluminum oxide was used as the abrasive. The groove was then treated with an HF dye which stained the *p*-material darker than the *n*-material. A microscope with a calibrated traveling crosshair was used to determine the width of the groove at the point where the lapping bar had just cut through the *p*-material and into the *n*-material. From the known width of the groove at this point and diameter of the cutting bar, simple geometry sufficed to determine the depth of the *p*-layer.

III. SPECTRAL-RESPONSE MEASUREMENTS

Fig. 2 summarizes the method of making the spectral-response measurements. A beam of essentially monochromatic light from a diffraction-grating monochromator *A* is incident upon position *B* where either a test cell or the calibrated standard cell is placed. For wavelengths longer than 0·6 μ a filter is placed at *C* to remove higher-order spectra. The light source for the monochromator is an incandescent bulb. The current produced by the cells is run through a 56-Ω load (an impedance low enough to be essentially a short circuit when compared to the high impedance of the dimly illuminated cell) and the voltage thus produced is measured by a d.c. microvoltmeter.

The standard cell is a solar cell made at Bell Laboratories several years ago and is encapsulated in a clear plastic. It was calibrated at Bell Laboratories against a vacuum thermocouple using a Bureau of Standards source.

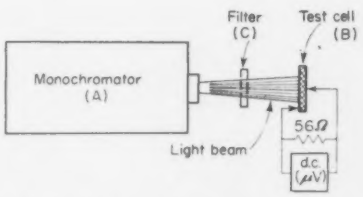


FIG. 2. Spectral-response measurement set-up.

The results of the spectral-response measurements on etched cells are summarized in Fig. 3. By use of the standard cell all measurements have been normalized to equal energy density incident upon the cell at each wavelength. The characteristics of the curves in Fig. 3 may be summarized as follows:

(1) For wavelengths of incident light shorter

than 0·75 μ the response increases for shallower junctions.

(2) For all but the deepest junction (about 5·0 μ) the response for wavelengths longer than 0·75 μ is virtually the same.

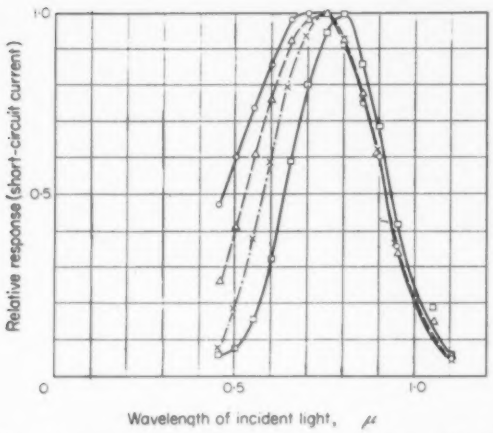


FIG. 3. Spectral-response curves for etched cells of varying junction depths. □ 288-min cell (5·0 μ); × 72-min cell (3·2 μ); △ 24-min cell (1·7 μ); ○ 6-min cell (0·6 μ).

(3) The wavelength of maximum response is 0·75 μ, except for the shallower junctions (below 1·0 μ), where the peak tends to flatten and broaden towards shorter wavelengths, and the deepest junction, where the peak moves towards longer wavelengths.

Curves similar to those in Fig. 3 were obtained from cells having lapped surfaces except that the lapped cells invariably had a poorer short-wavelength and better long-wavelength response than etched cells having the same diffusion time. These differences were most pronounced for cells with shallower junctions, and became less noticeable as the junction was moved farther away from the surface. Fig. 4 illustrates this trend.

IV. A SIMPLE THEORETICAL MODEL

This section will present a simple theoretical model of the solar cell to explain the spectral-response measurements.* This model is made under the following assumptions:

(1) One-dimensional geometry applies.

(2) No surface damage or irregularities are present.

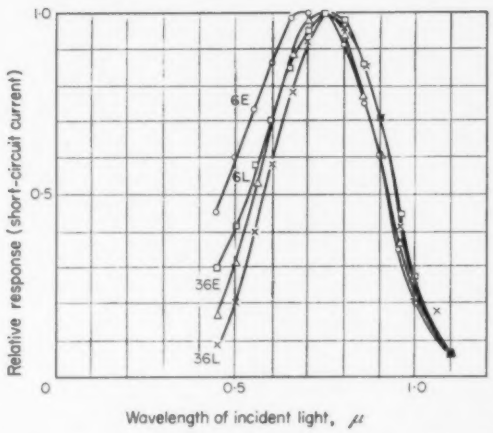


FIG. 4. Comparison of spectral-response curves for etched and lapped surfaces. ○ 6E (etched surface; 0·6 μ junction depth); □ 6L (lapped surface; 0·6 μ junction depth); × 36L (lapped surface; 2·0 μ junction depth); △ 36E (etched surface; 2·0 μ junction depth).

(3) The p-layer can be represented by a single lifetime.

(4) The fraction of incident photons reflected from the surface is independent of the wavelength of incident light.†

The densities discussed here are densities in excess of thermal equilibrium.

Fig. 5 indicates the geometry and co-ordinates assumed. For incident photons with energies in excess of the forbidden energy gap, the density of photons in the silicon varies as $N = N_0 \exp(-\alpha x)$ where N_0 is the surface density/in^2 and α is defined as the absorption coefficient.[6] Hole–electron pairs are created according to the negative of the rate of change of the photon density; thus the volume density‡ of hole–electron pairs created

* More general treatments of the theory of spectral response are given in Refs. 3–5.

† Although this assumption is not a valid one, it is felt that the effect of the dependence of reflection on wavelength could be accounted for by postulating slightly different values of L_n and L_p in the theoretical calculations to follow.

‡ For simplicity assume unit area.

by light photons varies as

$$N\,dx = N_0\alpha e^{-\alpha x}\,dx \qquad (1)$$

Minority electrons created at a distance d from the junction will have a fraction proportional to $\exp(-d/L)$ diffuse to the junction, where L is the diffusion length, and is equal to the square root of the product of the diffusion constant and lifetime. Thus

$$L = \sqrt{(D\tau)}$$

Now consider the hole–electron pairs created by photons in a volume element dx a distance x from the front surface. A fraction of the minority carriers created in this volume element will diffuse to the junction. The number which reach the junction will be proportional to the number created in the volume element, and the fraction which are able to diffuse the distance $|D-x|$ to the junction. Thus

density of minority carriers reaching the junction from a volume element dx located a distance x from the surface $\sim N_0\alpha \exp(-\alpha x) \exp -\left(\dfrac{|D-x|}{L}\right) dx \qquad (2)$

The number of minority carriers (electrons) reaching the junction due to creation of hole–electron pairs in the p-region is the integral of (2) from $x = 0$ to $x = D$, with $L = L_n$. Similarly, the number of minority carriers (holes) reaching the junction due to the creation of hole–electron pairs in the n-region is the integral of (2) from $x = D$ to $x = \infty$, with $L = L_p$. Thus

Total number of carriers crossing the p–n junction $\sim \dfrac{N_0\alpha}{\alpha - 1/L_n}\left[\exp -(D/L_n) - \exp(-D\alpha)\right] + \dfrac{N_0\alpha}{\alpha + 1/L_p}\exp(-D\alpha) \qquad (3)$

In the steady state the current through the cell is uniform everywhere. Hence it will be proportional to the number of carriers crossing the junction; since the right-hand side of (3) is proportional to this number, it is proportional to the total (or short-circuit) current.

N_0 is not a constant. For an equal-energy spectrum the density of photons is proportional to the wavelength of incident light, since the energy of an individual photon is proportional to frequency. Therefore, N_0 must itself be proportional to wavelength for an equal-energy spectrum. Thus

short-circuit current $\sim \dfrac{\lambda}{1 - 1/\alpha L_n}\left[\exp(-D/L_n) - \exp(-D\alpha)\right] + \dfrac{\lambda}{1 + 1/\alpha L_p}\exp(-D\alpha) \qquad (4)$

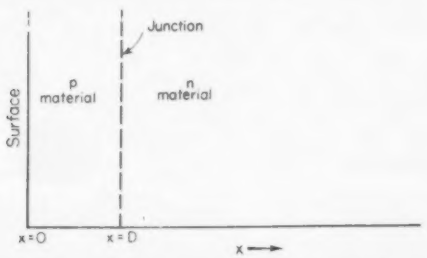


FIG. 5. Geometry assumed for theoretical model.

Since measurements of L_n and L_p are difficult, these parameters have been chosen for the best fit of the theoretical and measured curves. Values for α have been taken from the measurements of DASH and NEWMAN[7].

Fig. 6 indicates the good agreement that can be obtained between measured and calculated response curves using the values of L_n and L_p indicated on the figure. A 10-μ value for L_p corresponds to a lifetime of about 10^{-7} sec in the n-material, a quite reasonable estimate.

For etched-surface cells similar agreement can be obtained for all junction depths, although the values of L_n postulated must be made larger for deeper junctions and smaller for shallower junctions, while the value of L_p postulated remains essentially the same. One can conclude from this

that a considerable reduction in lifetime due to contamination exists near the surface, but, as the junction progresses deeper into the material, the

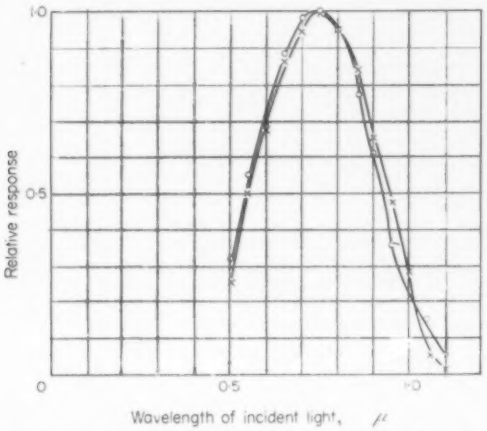


FIG. 6(a). Typical comparison between measured and calculated response curves for an etched surface. O Measured response for $D = 2{\cdot}0\,\mu$; × calculated response for $D = 2{\cdot}0\,\mu$; $L_n = 0{\cdot}5\,\mu$; $L_p = 10{\cdot}0\,\mu$.

effect of this surface contamination becomes less noticeable. Thus the average lifetime in the p-material will be longer for a deeper junction.

The effect of a lapped surface can be seen from Fig. 6(b). The assumption of $L_n = 0$ indicates

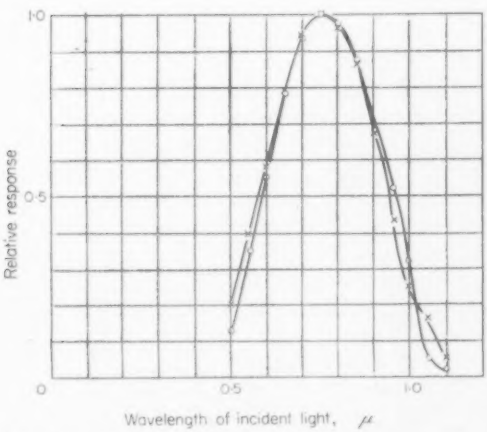


FIG. 6(b). Typical comparison between measured and calculated response curves for a lapped surface. O Measured response for $D = 2{\cdot}0\,\mu$; × calculated response for $D = 2{\cdot}0\,\mu$; $L_n = 0$; $L_p = 10{\cdot}0\,\mu$.

that the lifetime in the p-material has been reduced to the point where the contribution from the p-material is negligible. If the result of lapping the surface of the cells is to reduce the lifetime for a depth of several microns, it is evident that a phenomenon similar to the one observed in Fig. 4 would occur, since the effect of the lapped surface would become less important for deeper junctions.

The theoretical response curves are fairly sensitive to changes in the assumed values for L_n and L_p. A change of 2:1 in these parameters will change the theoretical relative response curve by as much as 20 per cent of the maximum response over appreciable portions of the curve.

Physically, the relative response curves fall off as λ approaches $1{\cdot}2\,\mu$ since α approaches 0 and the hole–electron pairs are created too deeply to be collected by the junction.

The relative response curves fall off faster than proportional to λ for $\lambda < 0{\cdot}5\,\mu$, since as λ approaches 0, α approaches infinity, and the hole–electron pairs are generated near the surface in low-lifetime material and recombine before reaching the junction.

If the lifetimes in the p- and n-material were infinite, the response curve would be a maximum at $\lambda = 1{\cdot}2\,\mu$, and would decrease linearly with decreasing λ until the response was zero at $\lambda = 0$ (for an equal-energy spectrum). Deviation from this straight-line response represents a loss mechanism which decreases the efficiency of the solar cell.

The effect of a reduction in lifetime in the vicinity of the surface due to surface damage or contamination is to reduce the short-wavelength response without effect on the response at longer wavelengths. This will also reduce the efficiency of the cell to a small extent.

The discussion of this section appears adequately to characterize the principal mechanism in determining the shape of the response curves.

Acknowledgements—The author is grateful to Dr. G. L. PEARSON who suggested the topic and supervised much of the work done. Dr. PEARSON also recommended the use of an evaporated aluminum dot as the p-layer contact. The author has benefited from a number of discussions with Dr. J. L. MOLL. Further, he would like to thank U. SORENSON and R. TAPERELL for their assist-

ance in the mechanics of making the solar cells, and W. SPITZER, of the Bell Telephone Laboratories, who calibrated the Standard Cell.

REFERENCES

1. M. PRINCE, *J. Appl. Phys.* **26,** 534 (1955).
2. D. CHAPIN, C. FULLER and G. PEARSON, *Bell Lab. Rec.* **33,** 241 (1955).
3. R. CUMMEROW, *Phys. Rev.* **95,** 16 (1954).
4. H. HARTEN and W. SCHULTZ, *Z. Phys.* **141,** 319 (1955).
5. W. PFANN and W. VAN ROOSBROECK, *J. Appl. Phys.* **25,** 1422 (1958).
6. T. MOSS, *Photoconductivity in the Elements* p. 16. Academic Press, New York (1952).
7. W. DASH and R. NEWMAN, *Phys. Rev.* **99,** 1151 (1955).

Solid-State Electronics Pergamon Press 1961. Vol. 2, No. 1, pp. 8–13. Printed in Great Britain

IMPURITY CONCENTRATIONS IN THE REGROWN REGION OF In–Ge ALLOYED JUNCTIONS

F. E. ROBERTS

Mullard Research Laboratories, Cross Oak Lane, Salfords, Nr. Redhill, Surrey

(*Received* 7 *June* 1960)

Abstract—The development of an infra-red microscope for routine carrier-density measurements on small semiconducting specimens is described. With its aid, a study has been made of In–Ge and In–Ga–Ge regrown-alloyed regions of a size normally used in transistors. Indium concentrations of $3–4 \times 10^{18}$ cm^{-3} have been found near p–n junctions alloyed at 615°C, in agreement with the established solidus curve at that temperature, but at 418°C the results would appear to favour a linear plot of log (distribution coefficient) versus $1/T$°K. In +0·62 at. per cent Ga pellets alloyed at 618°C yield acceptor concentrations of $1{\cdot}4 \times 10^{19}$ cm^{-3}. Diffusion lengths have been measured in these regrown regions by photoinjection methods and are found to be of the order of 5 μ.

Résumé—Cet article décrit le développement d'un microscope infrarouge pour mesurer ordinairement la densité des porteurs dans des petits spécimens de semiconducteurs. En l'employant, une étude a été faite des régions développées-alliées d'In–Ge et d'In–Ga–Ge, ayant les dimensions normales des transisteurs. Des concentrations d'indium de $3–4 \times 10^{18}$ cm^{-3} ont éte découvertes près des jonctions p–n alliées à 615°C en accord avec la courbe solide établie à cette température; mais à 418°C les résultats semblent favoriser une courbe linéaire du logarithme (coefficient de distribution) en fonction de $1/T$°K. Des pastilles de formation (In +0,62 pour cent atomique Ga) alliées à 618°C produisent des concentrations d'accepteurs de $1{,}4 \times 10^{19}$ cm^{-3}. Dans ces régions developpées par des méthodes d'injection photoélectrique les longueurs de diffusion ont éte mesurées et leurs valeurs se situent dans les environs de 5 μ.

Zusammenfassung—Die Entwicklung eines Infrarot- Mikroskops zur Messung der Trägerdichte auf kleinen Halbleiterproben wird beschrieben. Es wurde benutzt, um die rekristallisierten legierten Zonen von der bei Transistoren üblichen Grösse für In–Ge und In–Ga–Ge zu untersuchen. In der Nähe von bei 615°C legierten p–n-Übergängen ergaben sich Indiumkonzentrationen von $3–4 \times 10^{18}$ cm^{-3}, in Übereinstimmung mit den bekannten Festkörperkurven bei dieser Temperatur. Bei 418°C deuten die Ergebnisse eher eine lineare Abhängigkeit des Logarithmus des Verteilungskoeffizienten von $1/T$°K an. In Ga-Perlen mit +0,62 Prozent Atomen, bei 618°C legiert, ergeben sich Akzeptorkonzentrationen von $1{,}4 \times 10^{19}$ cm^{-3}. Die Diffusionslängen in diesen rekristallisierten Zonen wurden durch Photoinjektion gemessen. Sie sind von der Grössenordnung 5 μ.

I. INTRODUCTION

IT HAS long been known[1,2] that at wavelengths longer than the lattice absorption edge, semiconductors exhibit absorption of light due to free charge carriers, the absorption coefficient generally increasing with λ^2, and being proportional to carrier density. LEHOVIC[3] and GIBSON[4] have turned this property into useful photomodulator devices, while HARRICK[5] has used it for resistivity measurements and the study of injection, etc., in germanium.

This paper outlines the development of the infra-red microscope of AVERY[6] to the stage where it may be used for routine carrier-density measurements on small semiconducting specimens, and presents the results of a study of In–Ge, and In–Ga–Ge, recrystallized alloyed regions of a size normally used in transistors. Its application to the study of resistivity striations in pulled germanium crystals has already been described.[7]

2. THE INFRA-RED MICROSCOPE

The optical system of the infra-red microscope is shown in Fig. 1. Its characteristic feature is the

reflecting objective O, which enables focusing to be performed visually at E. The usual transmission formula has been used

$$t = (1-R)^2 \exp(-\bar{\alpha}d)/(1-R^2 \exp(-2\bar{\alpha}d) \quad (1)$$

where R, $\bar{\alpha}$, d are respectively the reflexion coefficient, a mean absorption coefficient and the thickness of the specimen, interference effects from multiple internal reflexions being neglected. In order to obtain reliable values of $\bar{\alpha}$ for a slightly wedge-shaped specimen, a germanium lens L was inserted so that the specimen S and the lead sulphide detector P were at conjugate foci. The diffusing glass screen G was found necessary to prevent saturation of a small area of the cell with intense light. Thus the wavelength range is limited by the absorption edge of the lens and by the sensitivity of the photocell to about 1·9–2·6 μ. This is wide enough to ignore interference effects yet (as can be shown) sufficiently monochromatic for the variation of α with wavelength to have no effect on the measurements.

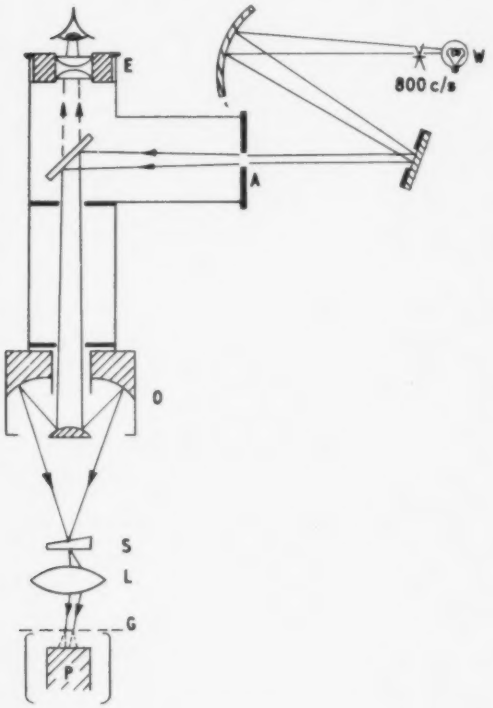


FIG. 1. Schematic diagram of the apparatus, showing the light path.

EATON[8] has examined the performance of the microscope in detail, and has shown that diffraction is the limiting factor in obtaining a small light spot. In these experiments, a nominal $20 \times 100\ \mu$ spot has been used, in which 90 per cent of the total energy lies within a strip 30 μ wide. The error in transmission of a slightly wedge-shaped specimen up to 1 or 2 mm thick has been measured on a trial specimen to be 5 per cent or less. Values of the product (αd) lying between 1 and 4 may usually be obtained with an overall uncertainty of 15–10 per cent in "Specimen In—Specimen Out" measurements; in relative measurements on the same specimen the uncertainty is 5–10 per cent.

The apparatus was tested and calibrated using a number of p-type Ge specimens with resistivities from $8{\cdot}2 \times 10^{-2}$–$8{\cdot}5 \times 10^{-4}$ Ω-cm. This range corresponds to $6{\cdot}8 \times 10^{16}$–$4{\cdot}6 \times 10^{19}$ holes or impurities per cm^3, according to the curve of TRUMBORE and TARTAGLIA[9] connecting resistivity with impurity concentrations in p-type germanium. A few specimens were checked by Hall measurements, with identical results. This work confirms the direct proportionality of $\bar{\alpha}$ with hole density:

$$\bar{\alpha} = 10^{-16}p$$

and agrees with known absorption cross-sections.[2]

3. EXPERIMENTAL METHOD

The object of the experiments was to measure impurity concentrations at the p–n junction as a function of alloying temperature, their variation through the recrystallized region, and hence the solid solubilities of the impurities. Also required was an estimate of the diffusion length in the regrown region.

The technique involved polishing the recrystallized material into a wedge, and rests on three assumptions, that: (i) the p–n junction is planar, (ii) impurity concentration depends only on distance from the junction, i.e. one-dimensional geometry is assumed, and (iii) equilibrium was maintained throughout the cooling cycle.

Fig. 2 shows a wedge of p^+-material lying on its transparent n-type base, and a beam of light focused on to the surface at the point P (x, z). The angle θ of the surface at P is given by

$\tan\theta = dz/dx$ and the absorption coefficient is $\alpha = \alpha(z)$. Then the transmission of light in the absence of multiple internal reflexions is

$$t \propto \exp\{-\int_0^z \alpha(z)\,dz\} =$$

$$\exp\{-\int_0^x \alpha(x)\tan\theta(x)\,dx\}$$

i.e.

$$\alpha(x) = -\frac{1}{\tan\theta(x)}\cdot\frac{d}{dx}(\log_e t) \qquad (2)$$

FIG. 2. Configuration of the wedge-shaped regrown region.

t and θ are measured quantities. Using

$$z = \int_0^x \tan\theta\,dx$$

α and, therefore, p are found in terms of z.

PFANN's[10] equation for normal one-dimensional freezing of a binary alloy when volume changes are ignored, is

$$C(T) = C_a/(1-g)^{1-k}$$

where $C(T)$, C_a are respectively the impurity concentrations in the liquid at temperature T and at the alloying temperature, as given by the liquidus curve.[11] g is the fraction of original liquid alloy which has solidified, and $k = C_s/C$ is the (constant) distribution coefficient, where C_s is the impurity concentration in the solid. In practice, k varies with T, but since for the impurities of interest it is very small, we may neglect it altogether and obtain

$$g/g_m = [1-C_a/C(T)]/[1-C_a/C_m] \qquad (3)$$

where $g = g_m$, $C = C_m$ at the eutectic or pseudo-eutectic, corresponding to the surface of the regrown region. The distance z from the junction is $z = (\text{thickness of regrown region}) \times g/g_m$. Equation (3) then connects z with temperature, and therefore, writing $p = C_s$[9], equations (2) and (3) together yield C_s versus T. In the case of the In+Ga metal pellets, it has been assumed that freezing was governed by the In–Ge liquidus. These pellets contained 0·62 at. per cent Ga.

Analysis of the freezing equation using an analytical expression for the liquidus, and of photo-injection of carriers, has been given in detail by GUDMUNSEN and MASERJIAN[12] who have allowed for differences in the atomic volumes of the constituent elements. The accuracy lost here by ignoring volume effects is small compared with the experimental uncertainty.

Diffusion-length experiments were made in exactly the same way as GUDMUNSEN's by injecting carriers from the surface with non-penetrating light, and picking up the signal across the junction as the light spot moved up the wedge. If L is the diffusion length, and $z \gg L$, the signal is proportional to $\exp(-z/L)$.

4. PREPARATION OF SPECIMENS

(111)-oriented 0·5 Ω-cm n-type germanium wafers were etched in alcohol etch, and the pellets in $HF:HNO_3:H_2O_2 = 2:1:1$ before insertion into the furnace, and heated at 10–20°/min in mixed gas. Subsequent preparation for optical measurements was as follows:

Excess metal over the alloyed region was removed in HCl, and the surface polished down at an angle of about 3°, until the recrystallized p^+-material lay as a wedge on the high-resistivity n-type base. The outline of the wedge was shown up by plating with black nickel as in Fig. 3, or by light etching. The flat on one side is where the junction meets the surface, and its straightness demonstrates the flatness of the junction. After a clean-up polish, a wire was soldered to the base, point-by-point measurements of surface angle were made, the underside was coated with $PbCl_2$[13] for maximum transmission at 2·2 μ, and finally the specimen was mounted with a tungsten point contacting the p^+-region for the diffusion-length measurements.

5. RESULTS

Table 1 presents the results of seven germanium

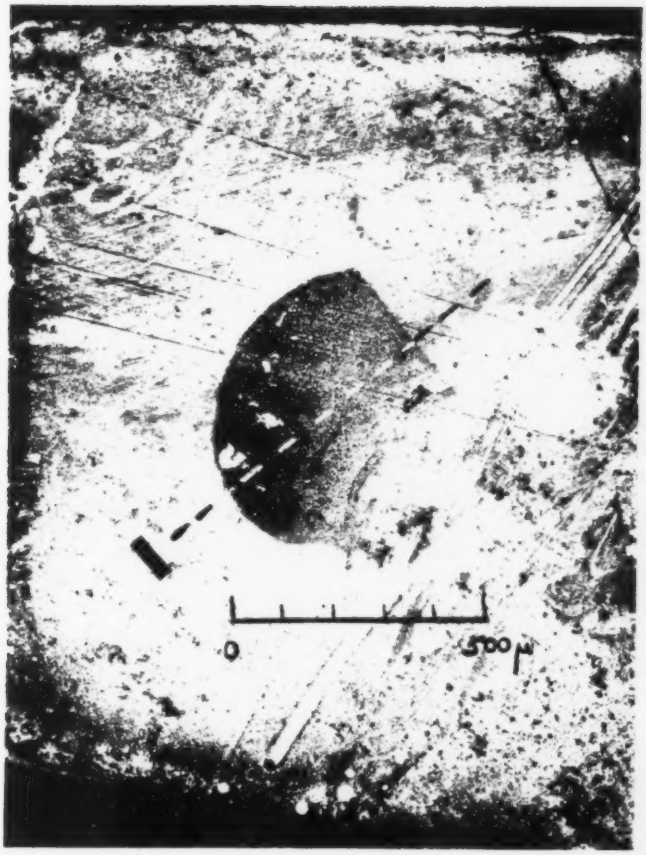


FIG. 3. Specimen III.1 after polishing and plating with black nickel, showing the track of the light spot during measurements.

[*facing p.* 10

specimens alloyed with indium, together with alloying data. Only the two specimens marked with an asterisk were good enough for the detailed analysis of Section 3, and their results are given as lines in Fig. 4, where impurity concentrations

The results of two specimens alloyed with In+Ga pellets are given in Table 2, and the calculated distribution coefficient of gallium is compared with that found by TRUMBORE *et al.*[19] for the binary Ge–Ga system. The calculation has been

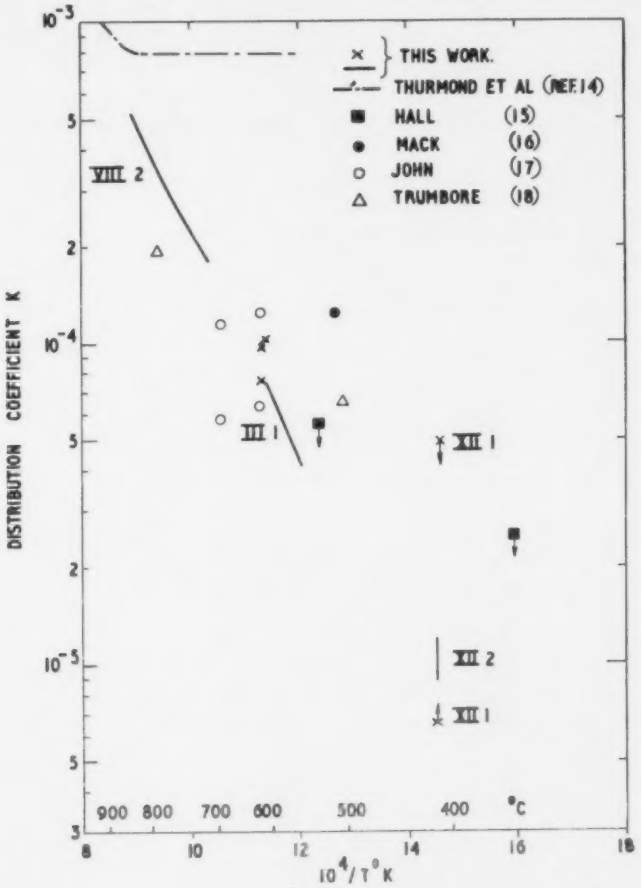


FIG. 4. The distribution coefficient of indium in germanium vs. $1/T°K$.

have been translated into distribution coefficients. Included for comparison are the results of THURMOND *et al.*[14], HALL[15], MACK[16], JOHN[17] and TRUMBORE[18]. Some difficulty was experienced in obtaining consistent results at the lower temperatures, but specimen XII.2 is believed to be fairly reliable. The uncertainty in XII.1 arises from several bad scratches.

made assuming that the amount of indium in solid solution is unchanged on adding gallium to the melt.

Except for specimen XII.1, the hole concentrations found here are thought to be correct to within a factor of two.

6. DISCUSSION

TRUMBORE[18] has recently reviewed the available

Table 1. *Results of alloying with indium*

Specimen no.	Alloying temperature (°C)	Alloying time (min)	Cooling rate (°C/min)	Doping at junction ($\times 10^{18}$cm^{-3})	Diffusion length (μ)
III.1*	610	7	9	3·2	5
III.2	610	7	9	4·4	3
IV.1	618	50	9·7	4·1	5
V.1	618	45	9·6	3·2	6
VIII.2*	881	3	~10	6	5
XII.1	418	120	1–3	0·3–2·5	—
XII.2	418	120	1–3	0·45–0·6	15

solubility data of impurities in germanium, and our results for indium at 600°C and above are in agreement with previous work. At lower temperatures, however, our results would seem to support HALL's[15] suggestion that $\log k \propto 1/T°\mathrm{K}$ and that his experimental points are maximum estimates, contrary to TRUMBORE's conclusion.

The calculated values of k in Table 2 for gallium in germanium are obviously correct only in order-of-magnitude, and, although comparison with the quoted values is interesting, no conclusion may be drawn without further work. If the increase in k is significant, it might be connected with the fact that the radius of a gallium atom in its liquid is about four-fifths that of indium, whereas it has nearly the same tetrahedral radius as germanium.

It is interesting to estimate the expected emitter efficiency of a junction from our results. WATKINS[20] quotes the efficiency $\gamma = 1/(1+\gamma')$, where

$$\gamma' = \frac{n_b L_b}{\mu_b} \frac{\mu_e}{p_e L_e}$$

and L is diffusion length or base width, μ is minority-carrier mobility. The subscripts b and e refer to the base and emitter sides respectively. Typical values for the n-type base are $n_b = 5\times10^{14}$ cm^{-3}, $\mu_b = 1800$ cm^2/V-sec, $L_b = 5\times10^{-3}$. For an indium-doped emitter alloyed at 600°C, $p_e = 3\times10^{18}$, $\mu_e \simeq 1000$ cm^2/V-sec (estimated from TRUMBORE and TARTAGLIA's data[9]), $L_e = 5\times10^{-4}$ cm, yielding $\gamma \simeq 0{\cdot}999$. DEB and DAW[21] have found $\gamma \sim 0{\cdot}994$ for the 0C70 transistor. For the 0C72, $p_e = 1{\cdot}4\times10^{19}$, we would have by comparison $\gamma \simeq 0{\cdot}9998$.

The diffusion lengths found in these regions, of the order of 5 μ, compare well with 4–12 μ diffusion lengths found in Si–Al regrown regions[12], and, from the formula $\tau = eL^2/\mu kT$, would imply a lifetime of about 3×10^{-8} sec.

7. CONCLUSION

The infra-red microscope of AVERY[6] has been successfully employed for the measurement of carrier concentrations in small semiconducting specimens. In particular, it has been used to study the doping levels and diffusion lengths in the re-crystallized regions of In and In+Ga alloyed junctions on germanium.

Solid solubilities of indium obtained in this way

Table 2. *Results of alloying with* In+0·62 *at. per cent* Ga

Specimen no.	Alloying temperature (°C)	Doping at junction ($\times 10^{19}$ cm^{-3})	Diffusion length (μ)	Distribution coefficient of Ga	
				Calculated	Ref. 18
V.4	618	1·4	—	0·04	0·015
VIII.3	881	1·3	2	0·15	0·035

Alloying times and cooling rates were as for specimens V.1 and VIII.2 respectively (Table 1).

are in agreement with other work at 600°C and above, but at lower temperatures would seem to favour HALL's suggestion of a linear plot of log k vs. $1/T$°K, rather than TRUMBORE's higher values.

Diffusion lengths of the order of 5 μ have been measured, and they compare well with GUDMUNSEN's results for Si–Al regrown regions.

Note added in proof: Results of two specimens alloyed at 404°C with larger quantities of In, and cooled at 2°/min, were: $7{\cdot}6\times10^{17}$ and $3{\cdot}5\times10^{17}$ holes/cm^3, giving $k = 1{\cdot}5\times10^{-5}$ and $6{\cdot}8\times10^{-6}$ respectively. Such low k-values agree with Specimens XII and with the recent work of M. A. LEE, *Solid-State Electron.* **1**, 194 (1960).

Acknowledgements—The author's thanks are particularly due to Dr. M. D. STURGE for suggesting the problem and for help with the apparatus; to Dr. F. A. KRÖGER for useful discussions; to Mr. R. PRATT for the Hall measurements; and to Mr. J. SEWELL, Mrs. K. BUCKTON and Miss B. SHOVE for help with polishing and the plating techniques.

REFERENCES

1. H. Y. FAN, *Rep. Progr. Phys.* **19**, 107 (1956).
2. T. S. MOSS, *Optical Properties of Semiconductors.* Butterworths, London (1959).
3. K. LEHOVIC, *Proc. I.R.E.* **40**, 1407 (1952).
4. A. F. GIBSON, *Proc. Phys. Soc.* **B66**, 588 (1953).
5. N. J. HARRICK, *Phys. Rev.* **101**, 491 (1956); *Solid-State Electron.* **1**, 234 (1960).
6. D. G. AVERY, *J. Sci. Instrum.* **34**, 16 (1957).
7. F. E. ROBERTS, *Solid-State Electron.* **1**, 93 (1960).
8. G. K. EATON, Unpublished work (1955).
9. F. A. TRUMBORE and A. A. TARTAGLIA, *J. Appl. Phys.* **29**, 1511 (1958).
10. W. G. PFANN, *J. Metals, N.Y.* **4**, 747 (1952).
11. C. D. THURMOND and M. KOWALCHIK, *Bell Syst. Tech. J.* **39**, 169 (1960).
12. R. A. GUDMUNSEN and J. MASERJIAN, *J. Appl. Phys.* **28**, 1308 (1957).
13. S. D. SMITH and T. S. MOSS, *J. Sci. Instrum.* **35**, 105 (1958).
14. C. D. THURMOND, F. A. TRUMBORE and M. KOWALCHIK, *J. Chem. Phys.* **25**, 799 (1956).
15. R. N. HALL, *J. Phys. Chem. Solids* **3**, 63 (1957).
16. G. MACK, *Z. Phys.* **152**, 26 (1958).
17. H. JOHN, *J. Electrochem. Soc.* **105**, 741 (1958).
18. F. A. TRUMBORE, *Bell Syst. Tech. J.* **39**, 205 (1960).
19. F. A. TRUMBORE, E. M. PORBANSKY and A. A. TARTAGLIA, *J. Phys. Chem. Solids* **11**, 239 (1959).
20. T. B. WATKINS, *The Junction Transistor and its Applications* (Edited by E. WOLFENDALE) p. 36. Heywood, London (1958).
21. S. DEB and A. N. DAW, *J. Electron. Control* **5**, 514 (1958).

Solid-State Electronics Pergamon Press 1961. Vol. 2, No. 1, pp. 14–17. Printed in Great Britain

A RAPID AND ACCURATE METHOD FOR MEASURING THE THICKNESS OF DIFFUSED LAYERS IN SILICON AND GERMANIUM

B. JANSEN

Philips Research Laboratories, N.V. Philips' Gloeilampenfabrieken Eindhoven, Netherlands

(*Received* 11 *July* 1960)

Abstract—Surface-layer thicknesses can be measured very practically by making use of the brittleness of silicon and germanium, thus eliminating cumbersome, precise grinding and polishing procedures. The germanium or silicon slice is broken in a special manner producing a rather flat cleavage surface, in which any disturbing fracture line is more or less perpendicular to the very sharp, long edges. In this surface the exposed p–n junctions are marked as very thin lines by one of the well-known methods (electrolytical, chemical). The distances between these lines or to the edge of the cleavage surface are measured under a high-power metallographic microscope with an eyepiece micrometer. Magnification factors mostly used are 400 or 600 times. Very thin layers of about 1 μ are still measurable. The method is used for diffused layers as well as for alloyed contacts or combinations as in the case of alloy diffusion.

Résumé—Les épaisseurs des couches de surface peuvent être mesurées d'une façon très pratique en faisant usage de la fragilité du silicium et du germanium, éliminant ainsi les procédés exacts et laborieux de polissage et de meulage. La tranche de germanium ou de silicium est fragmentée d'une façon spéciale produisant une surface de clivage plutôt plate, dans laquelle les lignes de fracture perturbatrices sont plus au moins perpendiculaires aux longues arêtes aigues. Dans cette surface, les jonctions p–n exposées sont marquées de lignes très fines par une des méthodes chimiques ou électrolytiques connues. Les distances entre ces lignes ou jusqu'à l'arête de la surface de clivage sont mesurées à l'aide d' un microscope métallographique de haute puissance ayant un oculaire-micromètre. Les facteurs de grandissement employés varient entre 400 et 600. De très fines couches d'environ 1 μ peuvent être mesurées. Cette méthode est employée aussi bien pour les couches diffusées que pour les contacts alliés ou pour des combinaisons comme dans le cas de diffusion d'alliage.

Zusammenfassung—Die Dicke von Oberflächenschichten lässt sich leicht messen, wenn man aus der Sprödigkeit von Silizium und Germanium Nutzen zieht. Man vermeidet dadurch umständliche Schleifund Polierprozesse. Die Germanium- oder Siliziumscheiben werden in bestimmter Weise zerbrochen, so dass eine flache Bruchfläche entsteht, in der störende Bruchlinien auf den scharfen, langen Kanten mehr oder weniger senkrecht stehen. Auf dieser Fläche werden die freiliegenden p–n-Übergänge durch eines der üblichen Verfahren (elektrolytisch, chemisch) als dünne Linien markiert. Die Abstände zwischen diesen Linien oder bis zu den Kanten der Bruchfläche werden unter einem hochempfindlichen metallographischen Mikroskop mit einem Okularmikrometer gemessen. Die Vergrösserung beträgt meist 400 bis 600. Auch sehr dünne Schichten bis etwa 1 μ sind noch messbar. Die Methode wird für diffundierte Schichten, für legierte Kontakte und für Kombinationen, wie im Fall der Legierungsdiffusion angewendet.

INTRODUCTION

CLEAVAGE surfaces have been found very useful in checking the diffusion of impurities into germanium and silicon and they have eliminated cumbersome precise grinding and polishing procedures. By proper breaking, the appearance of disturbing fracture lines parallel to the wanted p–n junctions is prevented. Corrections for the fact that the fracture surfaces are not exactly perpendicular to the layers are very small. The largest ever found was about 10 per cent.

PRINCIPLE

Because the edges of a cleavage surface running

through a piece of silicon or germanium are very sharp, as can be seen under a high-power metallographic microscope, they can be used as reference lines for a direct measurement of layer thicknesses. If the two following conditions are fulfilled, this measurement is also very exact. Firstly the position of the layer boundary in the cleavage surface must be sharply marked. Being mostly a *p–n* junction this is always possible by electroplating, chemical plating or etching. Secondly it is necessary that no fracture lines are parallel to the layer, otherwise they interfere with clear identification of the *p–n* junction, particularly in the case of very thin layers. It was found that, by proper breaking, these parallel fracture lines do not occur. The marking of the *p–n* junctions in this fracture surface can be done very sharply by using very diluted solutions and/or very short reaction times.

After proper adjustment of the finished specimen the distance between the edge of the cleavage surface and the *p–n* junction can easily be measured under a metallographic microscope with a magnification of about 400 times.

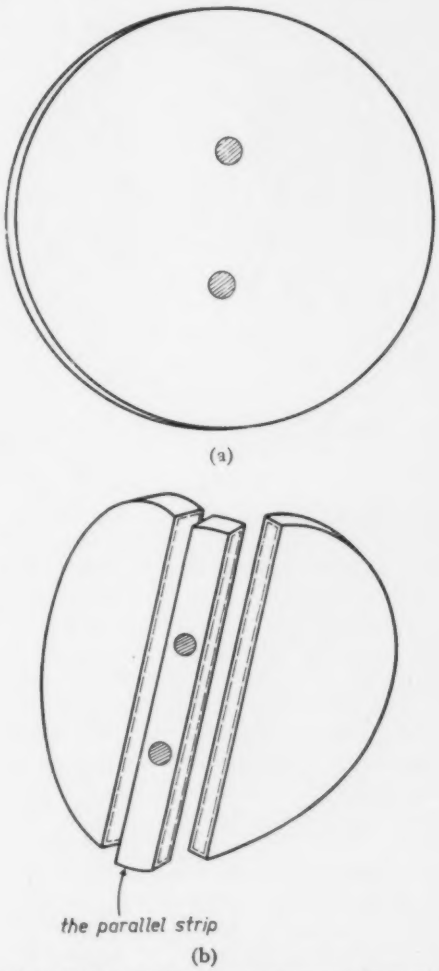
PREPARATION OF THE SAMPLES

The method requires the silicon or the germanium to have the form of slabs varying in thicknesses from 50 μ to 1000 μ. Protruding alloy dots are removed with a razor blade. There are no limitations to the dimensions. Only cubic pieces might give some trouble.

First of all a strip is broken out of the slab. For this purpose the slab is pinched in the jaw of a sliding gauge. In order to prevent the slab from cracking prematurely the jaws have to be covered with some elastic material, e.g. adhesive plastic tape. The slab is scratched with a fine diamond needle and then broken by exerting a light pressing force on the end of the slab. By doing this twice a narrow parallel strip is broken out of the slab [Fig. 1(a,b)].

The cleavage surfaces particularly fitted for exact measurements are obtained by breaking the strip as shown in Fig. 2, the diamond scratching being done at the lateral cleavage surface of the strip. At least a part of the cleavage surface so obtained will be rather flat without disturbing fracture lines (mostly the central part). Of course, one has to take into account the directions of fracture preference [the (111) planes]. Experience will, however, show the right direction of scratching.

After a flat cleavage surface has been obtained the *p–n* junctions exposed in it are marked.



(a)

(b)

FIG. 1(a). Original slab with two contacts, from which the protruding parts are removed. (b) Showing the parallel strip broken out.

In the case of *n*-type diffusion into *p*-type germanium this marking is done by copper-plating the *p*-type region of the *p–n* junction according to BOND and SMITS[1]. Fig. 3 shows such a marked *p–n* junction.

The plating of a very small specimen is facilitated

by the fact that it has several surfaces on which a whisker can be placed (Fig. 3). Firstly there is the cleavage surface in question, which is on top. Then there are two cleavage surfaces, together

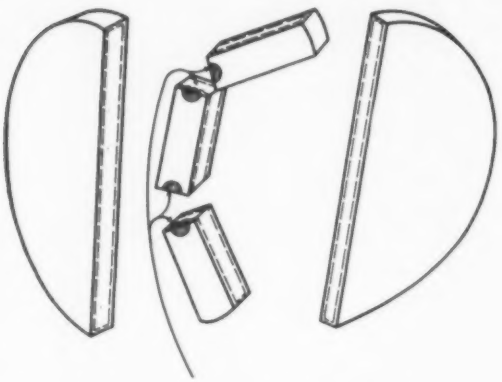


FIG. 2. Cleavage surfaces particularly fitted for exact measurements.

with the remaining parts of the top and bottom surface of the original slab, constituting the four vertical surfaces, seen in Fig. 4. Lastly there is the bottom surface which, as a rule, is also a cleavage surface. This latter surface, however, is used to fix the specimen to the adhesive layer of Scotch masking tape no. 202. This tape is stuck with some beeswax to a glass plate, the adhesive layer up. The adhesive layer does not contaminate the specimen as the wax does. After plating, the junction is seen as a line. In the case of high-ohmic *p*-type germanium it is necessary that the copper sulphate solution used to evaluate the *p*-region also makes contact with the 1·5 V probe, pressed against the *n*-type side of the sample*. If this contact is not made, the evaluation of the *p–n* junction is very poor as several parallel disturbing lines are formed.

In the case of *n*-type or *p*-type layers in silicon, obtained by a single or a double diffusion, the marking is possible by chemical plating or by etching. This simplifies the procedure quite considerably. In the case of etching, one uses a mixture of HF (40 per cent) and HNO_3 (65 per cent) (volume ratio 1:4) for only a very short time, mostly less than 1 sec (Figs. 5 and 6). Chemical plating is done with $Cu(NO_3)_2$ solutions according to WHORISKEY[2] or with AgF solutions according to ILES and COPPEN[3]. These solutions contain some HF and have been used more or

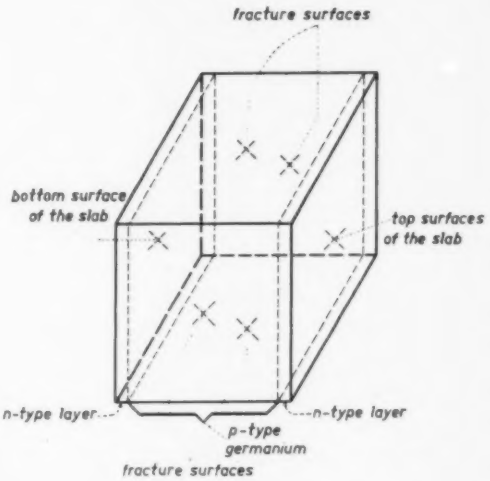


FIG. 4.

less diluted. This plating is best done under a low-power stereoscopic microscope and takes only a few seconds (Fig. 7).

Combinations of the marking methods are used in the case of alloy-diffused high-frequency germanium transistors. First of all the alloyed *p–n* junction is etched. Subsequently the diffused *n–p* junction is marked by electroplating of the interior *p*-type germanium with copper (Fig. 8).

MEASUREMENTS

After the *p–n* junctions have been marked, the sample is mounted in a little hill of beeswax, stuck on a glass plate. With the help of a stereoscopic microscope (6–80×) the cleavage surface is adjusted until it is perpendicular to the optical axis. In consequence of the curvature of the cleavage surface a very precise adjustment can be obtained only for part of the surface, but usually this is more than sufficient. After that, the thickness of the layer (*s*) is measured under a high-power metallographic microscope with the help of an

* See for this purpose: Fig. 4—Arrangement for Cu-plating the p-type side of a *p–n* junction in the article by BOND and SMITS[1].

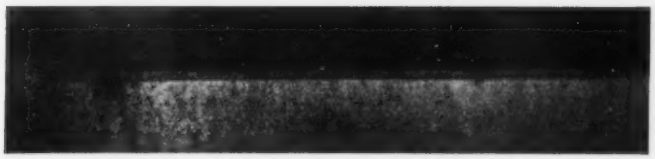


FIG. 3. *n*-Type layer in 1 Ω-cm *p*-type germanium: 600 ×; thickness: 2·5 μ.

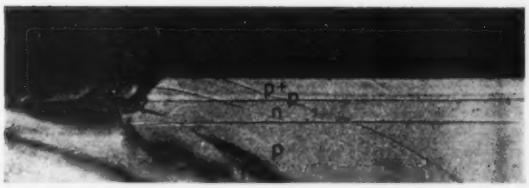


FIG. 5. *n*- And *p*-type layers in *p*-type silicon: 400 ×. Thickness of the outer *p*-type layer: 9 μ; thickness of the *n*-type layer: 9 μ; penetration depth of the alloyed contact about 6 μ.

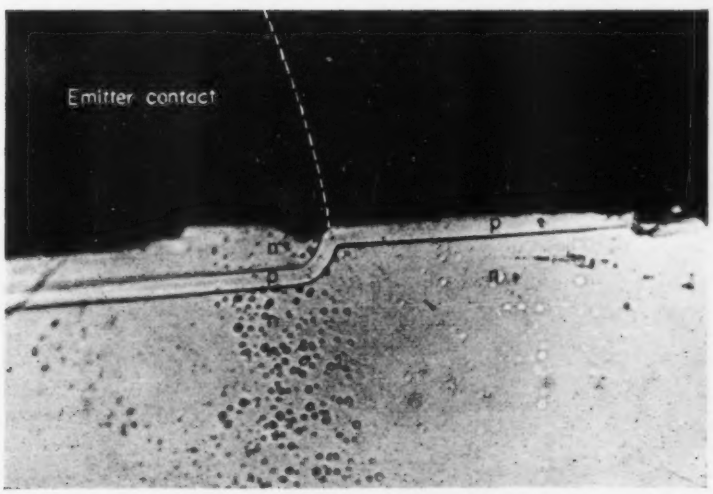


FIG. 6. Alloy-diffused *n*–*p*–*n* silicon transistor: 550 ×. Base width: 4·5 μ.

[*facing p.* 16

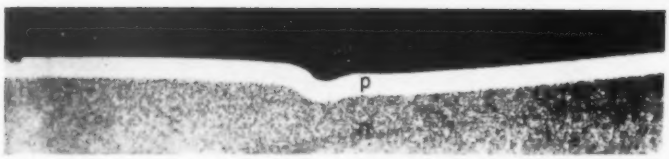


FIG. 7. *p*-Type layer of 9 μ in 1 Ω-cm *n*-type silicon: 375 ×.

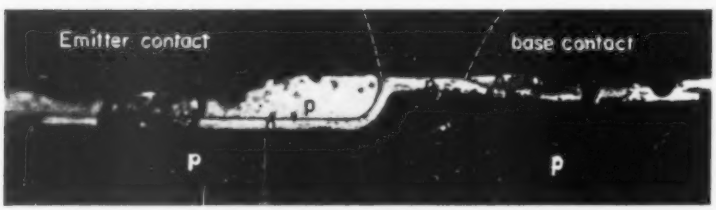


FIG. 8. Alloy-diffused *p–n–p* high-frequency germanium transistor: 600 ×. The *n*-type base layer is 3·5 μ thick.

eyepiece micrometer. The use of an object table with a ball-and-socket joint is recommended. Practice showed the small relative height differences in the used fracture surfaces to permit magnification factors of up to 1000. The magnification commonly used is 400 times. The agreement between samples taken out of one slab is excellent.

CONCLUSION

It is believed that the method described, using specially made cleavage surfaces, is of great help in developing and manufacturing semi-conductor devices both accurately and rapidly.

REFERENCES

1. W. L. Bond and F. M. Smits, *Bell Syst. Tech. J.* **35,** 1209 (1956).
2. P. T. Whoriskey, *J. Appl. Phys.* **29,** 867 (1958).
3. P. A. Iles and P. J. Coppen, *J. Appl. Phys.* **29,** 1514 (1958).

Solid-State Electronics Pergamon Press 1961. Vol. 2, No. 1, pp. 18–34. Printed in Great Britain

RECENT ADVANCES IN MICROMINIATURIZATION, RELIABLE COMPONENTS AND COOLING TECHNIQUES*†

G. W. A. DUMMER

Royal Radar Establishment, Malvern, Worcs.

(*Received* 27 *June* 1960)

Abstract—At the Royal Radar Establishment work has been done on electronic components, constructional techniques for equipments, including microminiaturization and cooling methods, for many years, and the article summarizes recent work in these fields. In microminiaturization techniques maximum attention is being given to the solid-circuit approach, with microcircuit on 2D being developed to obtain experience with film components. Circuits using all-silicon components have been constructed using transistor fabrication techniques and this technology is being expanded and the performance of solid-circuit components evaluated. Complete units are being built up similar to the Texas Instruments "Solid Circuit" units.

Most of the work on flat-plate or 2D circuits has been done on 0·6-in. square glass substrates, and resistors of evaporated nickel chromium using "in situ" copper masks also evaporated have been successfully made and evaluated.

Evaporation of dielectric materials for capacitors has not been so successful and a comprehensive programme of research into the stresses and strains developed in thin films, measurement techniques for accurate deposition and the study of initial stages in the nucleation and growth of thin films using radioactive tracer techniques, are all being carried out. Films have been made inside an electron microscope of the actual growth and nucleation processes of thin films. Ferromagnetic films have been made and the work is described. Transistors for microminiaturization work have been fabricated, initially following the work of Diamond Ordnance Fuse Laboratories in U.S.A. but later using hermetically sealed constructions 0·125 in. diameter ×0·040 in. deep. Microcircuits using evaporated components are also being assembled in tube envelopes as a means of protection and sealing.

Reliability problems have been studied at R.R.E. for many years and many analyses of faults occurring in equipments are given in the paper. A special range of "reliable" or "special quality" tubes and components is being developed and the stages in this programme are given. Results of a mass-testing evaluation of fixed resistors are detailed and some recent component developments are described, particularly some new miniature relays.

The life and performance of components at excessively high temperatures is summarized and work on liquid-cooling techniques is described in some detail, including proposed cooling and construction of microminiature assemblies. U.K. military experience with wrapped joints is described and the development of new miniature wrapped joints for use in connecting microminiature assemblies is given.

Résumé—Cet article résume les récents travaux et recherches entrepris par le Royal Radar Establishment depuis plusieurs années. Ces travaux comprennent les matériaux électroniques et les méthodes de construction de l'appareillage comprenant la microminiaturisation et les procédés de refroidissement. En ce qui concerne les méthodes de microminiaturisation, l'intérêt a été surtout porté aux procédés de circuit solide en développant le microcircuit sur 2D pour acquérir de l'expérience avec des matériaux à pellicule. Des circuits employant des matériaux entièrement au Si ont été construits en utilisant les méthodes de constructions des transistors; cette technologie est entrain d'être étendue et le rendement des matériaux des "circuits solides" calculé. Des unités complètes, analogues à celles de circuits solides de la Texas Instrument, sont entrain d'être construites.

* This paper was originally published in the Proceedings of the 1960 Electronic Components Symposium, Washington, sponsored by the I.R.E. and A.I.E.E.

† The British Crown Copyright is reserved on all photographs and illustrations used in this paper. They are reproduced here by permission of The Controller, Her Britannic Majesty's Stationery Office.

La plupart des travaux entrepris sur les circuits à 2D ou à surface plate ont été faits sur des verres carrés de 2,32 cm^2; des résistances de nickel chromé évaporé, utilisant des masques en cuivre "in situ" évaporés aussi, ont été fabriquées avec succès et leur valeur calculée.

L'évaporation des matériaux diélectriques pour les condensateurs n'a pas été aussi réussie. Un vaste programme de recherches comprenant: (1) une étude des fatigues et des déformations qui se developpent dans les pellicules fines, (2) de nouveaux procédés de mesures exactes du dépôt, (3) une étude des étapes initiales de la "nucléation" et du développement des pellicules fines employant des méthodes de tracement radioactif a été enterpris. En employant un microscope électronique, on a filmé le développement actuel et les transformations de "nucléation" des pellicules fines. Des mesures ferromagnétiques ont été prises et on décrit le travail entrepris. En suivant au début les travaux des laboratoires "Diamond Ordnance fuse" mais en employant plus tard les constructions hermétiquement scellées de dimensions 0,317 cm de diamètre et de 0,102 cm de profondeur, des transistors pour la microminiaturisation ont été fabriqués. Des microcircuits employant des matériaux évaporés ont été enveloppés dans des tubes pour les protéger et les sceller.

Les problèmes concernant la sûreté de fonctionnement ont été étudiés depuis plusieurs années au R.R.E. et plusieurs analyses de fautes qui se produises sont décrites. Une gamme spéciale de tubes et matériaux sûrs ou de qualité exceptionnelle est entrain d'être developpée et les étapes de ce programme sont ensuite décrites. Les resultats issus d'épreuves en masse de résistances fixes sont expliqués et on décrit quelques récents développements de matériaux, en particulier de nouveaux relais miniaturisés.

La longévité et le rendement des matériaux à des temperatures très elevées sont resumés et le travail entrepris sur les méthodes de refroidissement liquide est décrit en détail en y incluant les assemblages de construction en miniature. L'expérience militaire du Royaume-Uni en ce qui concerne les jointes enroulées est décrite ainsi que le développement des nouvelles jointes enroulées en miniature qui pourraient être utilisées pour les connections d'assemblages microminiaturisés.

Zusammenfassung—Im Royal Radar Establishment werden seit Jahren Untersuchungen über elektronische Bauteile und Konstruktionsverfahren für Geräte ausgeführt. Zu den letzteren gehört die Herstellung von Kleinstgeräten und Kühlungsmethoden. Die neueren Ergebnisse dieser Studien werden hier erörtert. In der Herstellung von Miniaturbauteilen beschäftigt man sich hauptsächlich mit Festkörperkreisen, und ein Mikrokreis auf 2D wird entwickelt, um Auflagekomponenten zu prüfen. Unter Verwendung der Transistortechnik wurden Kreise nur aus Siliziumkomponenten hergestellt, und dieses Verfahren wird ausgebaut, und die Leistung der Festkörperkreise wird ausgewertet. Die hergestellten Geräte gleichen den „Festkörperkreis-Geräten" der Texas Instruments.

Die Arbeit mit flachen oder 2D-Kreisen erfolgt meist auf 1,5 cm Quadraten aus Glassubstraten, und Widerstände aus aufgedampftem Chromnickel mit „in sitü" aufgedampften Kupfermasken wurden erfolgreich hergestellt und ausgewertet.

Die Aufdampfung von Dielektrika für Kondensatoren war weniger erfolgreich, und man plant ein umfangreiches Programm von Untersuchungen der Spannungen und Drucke in dünnen Auflagen, Messverfahren zur genauen Ablagerung und die Beobachtung der Anfangsstadien der Kernbildung und des Wachstums dünner Auflagen unter Verwendung radioaktiver Indikatoren. Mittels eines Elektronenmikroskops wurden Filme des Wachstums- und Kernbildungsprozesses von dünnen Auflagen hergestellt. Eisenmagnetischer Auflagen hergestellt wurden, und die Arbeit beschreibt ist. Transistoren für Kleingeräte fabriziert wurden zuerst nach dem Vorbild der Diamond Ordnance Laboratorien in USA, später unter Verwendung hermetisch verschlossener Kapseln von 3 mm Durchmesser und 1 mm Tiefe. Mikrokreise mit aufgedampften Komponenten werden ebenfalls in Röhren eingeschlossen, um sie zu schützen und abzuschliessen.

Die Zuverlässigkeit von Geräten ist ebenfalls seit Jahren im RRE geprüft worden, und der Artikel gibt Analysen von Defekten in Geräten. Eine Sonderreihe von Röhren und Komponenten, die als „zuverlässig" oder „Spezialqualität" markiert werden sollen, ist in der Entwicklung, und der Fortschritt dieses Programms wird beschrieben. Es folgen Ergebnisse der Auswertung einer Massenprüfung von Widerständen und die Beschreibung neuer Bauteile, insbesondere von Miniaturrelays.

Ferner findet sich ein Bericht über die Lebensdauer und Leistung von Bauteilen bei sehr hohen Temperaturen und von Untersuchungen über Kühlverfahren mit Flüssigkeiten, einschliesslich von Plänen für die Kühlung und Konstruktion vpm Kleinstaggregaten. Die Erfahrungen in der britischen Armee mit umwickelten Kontakten werden dargestellt, und man plant die Entwicklung neuer umwickelter Kontakte in Miniaturausmassen zur Verwendung in Kleinstaggregaten.

MICROMINIATURIZATION

At the Royal Radar Establishment, work on thin film components has been done since about 1946, and in May 1952 at the Electronic Components Symposium, Washington, the author put forward the following:

> "At this stage I would like to take a peep into the future. With the advent of the transistor and the work in semiconductors generally, it seems now possible to envisage electronic equipment in a solid block with no connecting wires. The block may consist of layers of insulating, conducting, rectifying and amplifying materials, the electrical functions being connected directly by cutting out areas of the various layers."

Since that date, interest in microminiaturization has grown, not necessarily for its potentialities as regards size reduction but because of the possible ultimate increase in reliability.

Of the four main current systems of microminiaturization, the micro-module, the micro-circuit (or 2D), the solid circuit and the subminiature approach, the last three are being worked on in Britain. Maximum attention is being given to the solid-circuit approach, with the micro-circuit or flat-plate circuit as the medium on which to gain reliability experience. Microminiature film components are also being developed for use inside sub-miniature metal-tube envelopes as the associated tubes resist nuclear radiation better than semiconductors. This technique also solves the film-protection problem and units so constructed should be able to withstand most known military environments.

Solid circuits

Although work on solid circuits was begun in the United Kingdom in 1956, it has not been possible to carry it out on the scale possible in the U.S.A., and more theoretical work has been done than practical work. The objective is the same as that envisaged in the U.S.A., i.e. the fabrication of working circuits using doping, shaping and other techniques in single crystals of silicon.

The following basic techniques are being studied in relation to solid-circuit research and development:

(1) Controlled diffusion of phosphorus and boron into silicon.

(2) Study of the depth of the diffused junctions by selective etching or staining.

(3) Study of the diffusion of impurity elements through oxide films on the surface of the silicon.

(4) Establishment of the technique of thermo-compression bonds.

(5) Study of double-diffusion methods for transistor fabrications.

(6) Control of the resistivity of paths in a silicon wafer by sandblasting slots in specific regions.

(7) Theoretical study of the construction of solid circuits.

(8) Simulation of solid circuits using individual components to determine tolerance limits.

As a first exercise in diffusion techniques, a simple distributed network has been made up as shown in Fig. 1, using *p*-type silicon bar, with a diffused *n*-type layer. Provided that the external

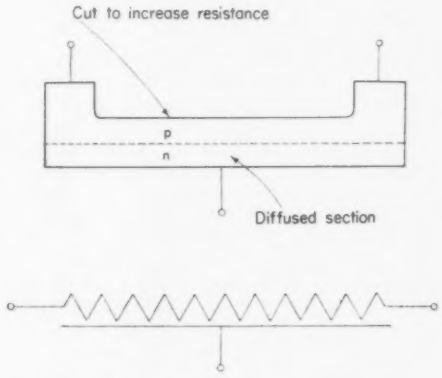


Fig. 1. Simple distributed network in silicon.

circuit is arranged so that the *p*–*n* junction is always reverse biased, this element is capable of a phase shift of greater than 180°, as was demonstrated by Texas Instruments. Twenty-seven decibels of power gain was necessary to use this element as a phase-shift oscillator. The oscillator frequency was 0·5 Mc/s.

A simulated solid circuit using all-silicon components has been constructed for tolerance tests. A simple all-silicon multivibrator was constructed using the circuit shown in Fig. 2. The four resistors were made from an H-shaped piece of silicon suitably dimensioned and the capacitors

were p–n junctions. Silicon transistors were used. The circuit functioned satisfactorily.

A paper study on the possibility of constructing a typical multivibrator circuit was made, in which the general principle adopted was that the whole

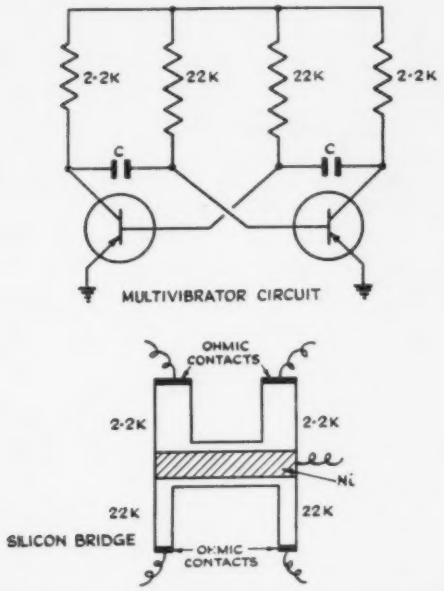


FIG. 2. Silicon resistors for multivibrator circuit.

circuit should be formed from the silicon wafer, the only external elements being insulating and conducting paths and the leads.

It is possible that, by a suitable re-design of the circuit to suit the techniques of solid circuitry, a saving of components would result. However, this paper exercise on a circuit specifically designed for conventional components illustrates the type of reasoning which must be adopted by the solid-circuit-device engineer. The techniques envisaged in this construction would be chiefly diffusion, oxide masking, evaporation of gold and gold compression bonding. The broad shaping of the specimen might be done by a suitable ultrasonic tool, and the final adjustment of resistance values by sand blasting through a suitable mask. The circuit, symbolic component layout and each side of the silicon-block layout is shown in Fig. 3.

This particular design is in terms of a 0·175 mm wafer of 25 Ω/cm p-type material. The resistive portions are largely made from strips 1·5 mm wide. The resistors of value 1, 3, 10 and 12 kΩ require lengths of 1·1, 3·2, 10·5 and 12·6 mm respectively. The four 3 kΩ resistors replace two 1·5 kΩ resistors in the original circuit, because of the particular method adopted of solving the cross-over problem associated with the linkage of the two halves of the multivibrator circuit. Silicon of this resistivity gives about 3000 pF/cm² in a step junction with low resistivity n-type material at 3 V reverse bias. The 100 pF capacitors are accordingly 1·5 × 2·2 mm in area, and the 200 pF capacitors 2·5 × 2·6 mm. Although the capacitance is voltage-dependent, a time constant is associated with the parallel combination of diode and resistor which can be adjusted to the required value of 1·2 μsec. A double-diffused structure is envisaged for the transistor. An alloy diffusion type can also be used, coupled with considerable reduction in the wafer thickness by jet etching.

The problem of connecting in stack form circuits or piece parts of circuits is being investigated, using the thermo-compression bond. If two circuits were designed to occupy two silicon slices as shown in Fig. 4(a) and (b) in which the contacts were raised and oriented (one to be a mirror image of the other), then, by gold plating the shaded areas to produce ohmic low-resistance contacts to the silicon and super-imposing (a) on (b) followed by the application of heat and pressure, a bond might be achieved and the two circuits connected as in Fig. 4(c).

Whilst this study demonstrated the feasibility of this construction, it was felt that it was too complex for fabrication with a low reject rate. Simpler circuits (e.g. NOR) are therefore being used, as two such units, joined as shown in Fig. 4, would form a multivibrator. It is not possible, of course, to match the effort available in the U.S.A., but the importance of this development is recognized in the United Kingdom and will be accelerated as effort becomes available.

Microcircuits

Most of the work at R.R.E. has been on the micro-circuit approach, with the aim of using evaporation techniques throughout as far as possible. It was felt that, although units could be made on "hearing aid" techniques using inserted ceramic capacitors, etc., it would be better to

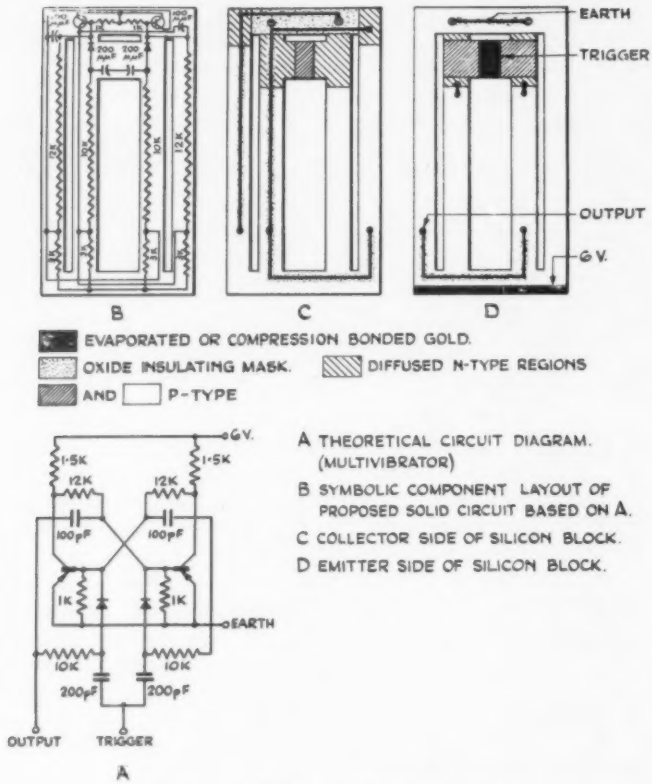


FIG. 3. Layout of silicon multivibrator block.

make a study of the full possibilities of evaporation, because the techniques of evaporation can be applied equally well to conductive, resistive, dielectric and magnetic films.

The basic stages in the R.R.E. process can be summarized as follows:

(1) Glass substrate prepared $0{\cdot}6 \times 0{\cdot}6 \times 0{\cdot}1$ in.
(2) Holes ultrasonically drilled for transistors.
(3) Copper evaporated over one surface.
(4) Resistor negative pattern and wiring etched in copper.
(5) Nichrome evaporated and copper etched away.
(6) Capacitor base electrodes evaporated (gold).
(7) Dielectric evaporated.
(8) Flat transistors inserted.
(9) Capacitor top electrode and all wiring evaporated in gold.
(10) Final encapsulation or sealing.

Masks are made of copper sheet, using normal photo-mechanical circuit processing etched right through. Wiring is first evaporated in chromium which keys to the glass (Stage 4) and over-evaporated in gold for conduction (Stage 9).

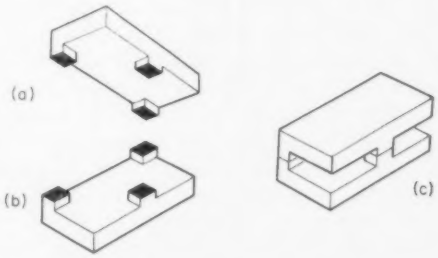


FIG. 4. Thermo-compression bonding of units.

Fixed resistors

Following the work of ALDERSON and ASHWORTH[1], nichrome resistors were evaporated and, after considerable experimental work, nichrome film resistors are now made at R.R.E. in the following way. The glass substrate is washed and cleaned, and a thin layer of chromium (~5Å) is deposited by vacuum evaporation, which acts as a bond for a thicker evaporated layer of copper (~1000 Å). The copper is coated with a light-sensitive resist and exposed to ultra-violet light through the required negative. The non-exposed parts of the resist are removed during subsequent development and etched in ferric chloride. This leaves a copper negative on the surface of the substrate. After suitable cleaning steps, it is put in a vacuum evaporator and nickel chromium is deposited by evaporation on to the substrate, which is maintained at 350°C. A controlled square is used to monitor the amount of nickel chromium being deposited. A value of 250 Ω/square is aimed at, which corresponds to a film thickness of 75 Å. It is annealed for half an hour before removal from the evaporator. The substrate is then removed and etched in strong ferric chloride which undercuts and removes the nichrome-covered copper regions, leaving the desired nichrome pattern on the glass. After suitable washing and drying the resistor is ready for use in subsequent stages. The stages in this process are shown in Fig. 5.

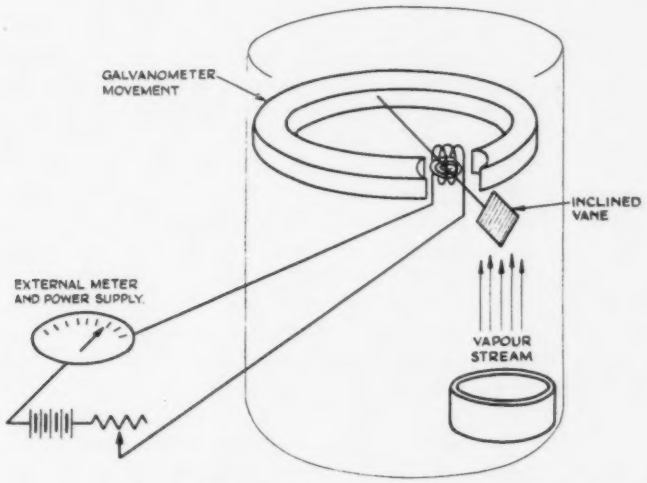


FIG. 6. Method of determining evaporation rate in vacuum.

Dielectrics

Evaporation of dielectrics for thin-film capacitors was, and still is, a difficult problem. Initial experimental work commenced with the use of lens-blooming materials because of their ease of evaporation. The principal materials were zinc sulphide, magnesium fluoride and silicon monoxide, all of which have fairly low permittivity. Early studies with these materials showed that very much more must be known about the nature of the films deposited, the control of the deposition processes, and the effects of impurities and substrates, before evaporation of materials with higher permittivity could be successfully undertaken. Consequently a number of research contracts have been initiated in order to increase available knowledge on these subjects.

It was soon realized that evaporated dielectric films can possess considerable internal stress. In the case of magnesium fluoride films, those thicker than 5000 Å were liable to craze spontaneously. Stresses can be either compressive or tensile. Factors influencing the stress include deposition

rate, substrate temperature and film thickness. This work has necessitated the development of methods for measuring stress by observing the flexure of a thin substrate during evaporation, and for precise control of evaporation rates. For this latter purpose an elegant flowmeter has been constructed, shown in Fig. 6. This consists of an obliquely mounted vane in the vapour stream which is deflected by molecular impact. It is hung from a galvanometer winding and it can be restored to its original position by a suitable current which is a measure of the deflecting force and hence the rate of flow of the vapour stress. This device could be used for automatic control. The stresses in the films have been related to the electrical properties of the films and to their mechanical stability. Results show that silicon monoxide and zinc sulphide have compressive stresses and are hence easier to deposit as thick films. Final conclusions have not yet been reached on this and work is continuing.

A start has also been made on the study of initial stages of nucleation and growth of evaporated films. At present this work is confined to thin metal films, but it is expected that the results from these will be applicable to dielectric films. Methods are being developed for evaporating radioactive materials, using the degree of radioactivity as a measure of film thickness and relating this to the onset of conductivity. Combined with this, films are to be evaporated inside an electron microscope and the initial stages of nucleation and growth observed. Also being studied is the effect of substrate cleaning on the nucleation and growth of films, again using radioactive carbon compounds as cleaning agents, measurements being taken after cleaning of residual surface activity to indicate the degree of contamination.

Work has also commenced on adhesion of evaporated films to various substrates, as this is a significant factor in their mechanical deterioration, as instanced by crazing and moisture absorption. Following work already done on the adhesion of metal films, the techniques developed are being extended to evaporated dielectric films. The research includes measurement of binding energies which it is hoped to correlate to adhesion.

Work at R.R.E. has included the initial evaluation of lens-blooming materials as dielectrics, the study of their leakage characteristics, voltage breakdown and susceptibility to atmospheric moisture. Work is in progress on the control of films involving two components evaporated from separate boats, with a view to the evaporation of materials with higher dielectric constant. At present studies are confined to boric oxide/silicon monoxide mixtures, but it is hoped shortly to extend these to lead oxide/silicon monoxide mixtures, and then to higher K materials including titanates.

Thin ferromagnetic films

Thin films of ferromagnetic materials about 1000 Å in thickness may be prepared by evaporation in a vacuum on to a suitable substrate, (carefully cleaned and heated during deposition to 300°C). Such films exhibit magnetic properties of a similar order to those of bulk material. Owing to the lack of substance of the films, their properties prove very sensitive to stress, unless a near-zero magnetostriction alloy such as 80/20 nickel–iron is used. If such films are deposited on an amorphous substrate so that a randomly oriented polycrystalline film is obtained, magnetocrystalline anisotropic effects are averaged out in all directions in the plane of the film. If the angle of incidence of the vapour beam to the surface during evaporation is kept below 10° and the substrate is smooth, other unwanted effects are made insignificant and a controlled preferred axis of magnetization in the plane of the film can be induced by applying a magnetic field parallel to the substrate during deposition.

Films prepared in this way exhibit a rectangular BH loop along the preferred axis and can be used as memory elements in a random access store which relies on coincident current operation. They behave in a manner similar to ferrite cores but if the switching field is applied at 10° or 20° to the preferred axis, very rapid switching from one magnetic state to the other can be obtained. Speeds of 50 mμsec in fields of twice the coercive force H_c have been measured. Unfortunately the coercive force of about 2 oersteds is high compared with ferrite cores and requires currents of an ampere or so in single turns or else multiturn windings. Such windings need not thread the film. They can be laid directly on top of an insulating layer on the magnetic film.

A further advantage in using such magnetic films

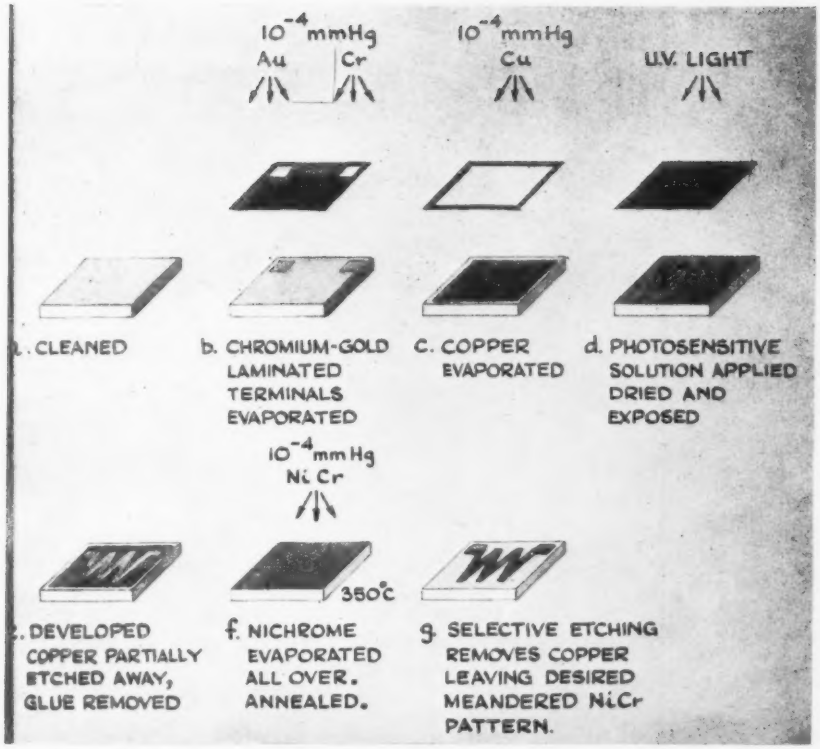


FIG. 5. Stages in the production of high-definition film resistors with terminals.

[facing p. 24

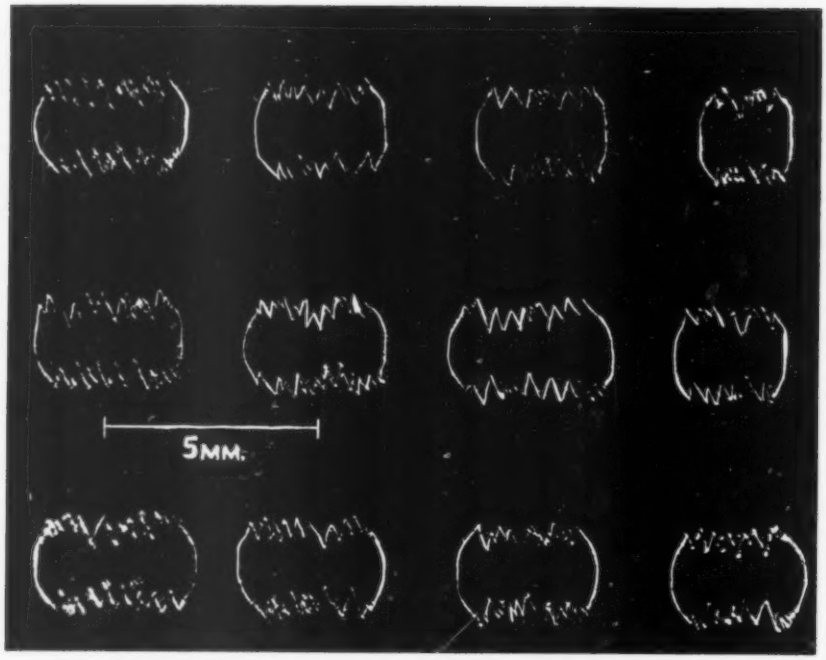


FIG. 7. Domains of reversed magnetization in Ni–Fe film.

is that they need not be broken up into discrete elements. A single large sheet can be used and the magnetization changes only in that area subjected to the reversing field. Fig. 7 shows such a sheet subjected to a reversing field $H>H_c$ at the coincidence of the two sets of drive wires. In these areas the magnetization has been reversed and the boundaries between regions of opposite magnetization have been shown up with colloidal magnetite. The regions between these spots were subjected to a reversing field of $H/2>H_c$ as would be the case in a coincident current type of store. The size of the discrete elements is determined by the demagnetizing effect of their edges and can be reduced considerably over that shown if the films are made thinner, although this of course reduces their output (which is normally a few millivolts per turn). By applying drive fields at right angles to the preferred axis of magnetization, speeds of switching of one or two millimicroseconds can occur. The return to one or other of the preferred directions under a small bias field is as fast. But here miniaturization is limited by the design requirements of the associated circuits to operate at these speeds.

Some thoughts are being given to the problem of similar evaporation techniques or ferrite deposition for radio-frequency inductors, but little work has yet been done.

Reliability evaluation of micro-circuits

Following the experimental R.R.E. work described previously, contracts have been placed with industry to manufacture large quantities of flat-plate circuits. The initial use will be for missile computers. Ten thousand resistors and 10,000 capacitors, together with several thousand transistors are to be manufactured in order to assess the reliability of the techniques before the final equipments are constructed.

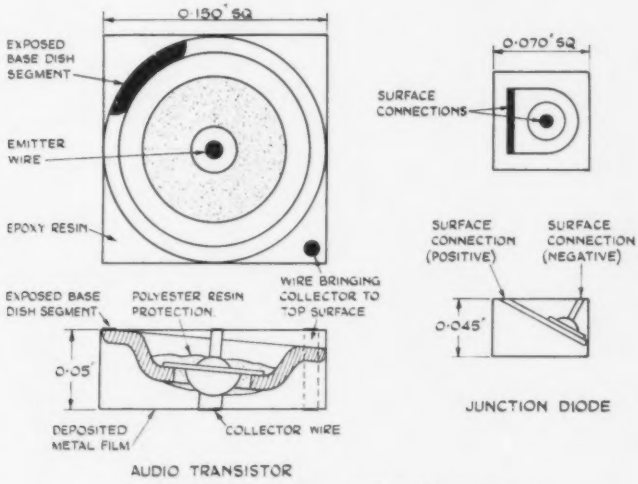


FIG. 8. Transistor and diode fabrication.

Transistors

Transistors are being developed in circular shape fully sealed by ceramic or glass–metal techniques and are to be in containers 0·125 in. in diameter and approximately 0·04 in. deep. The number of leads or contacts on the upper surface has not yet been decided, but will probably be more than three in order to assist circuit layouts. The initial work at R.R.E. has followed the lines of Diamond Ordnance Fuse Laboratories' pioneering work, and is illustrated in Fig. 8.

Tube assemblies

Tube assemblies are being developed in which individual assemblies of film-type components with suitable interconnexions are arranged in three-dimensional form in a sub-miniature tube envelope such as that shown in Fig. 9. They are used in conjunction with sub-miniature tubes in

which great care has been taken to provide hermetic seals, and also in the control of materials used in the techniques, in order to provide high reliability under severe military conditions. The

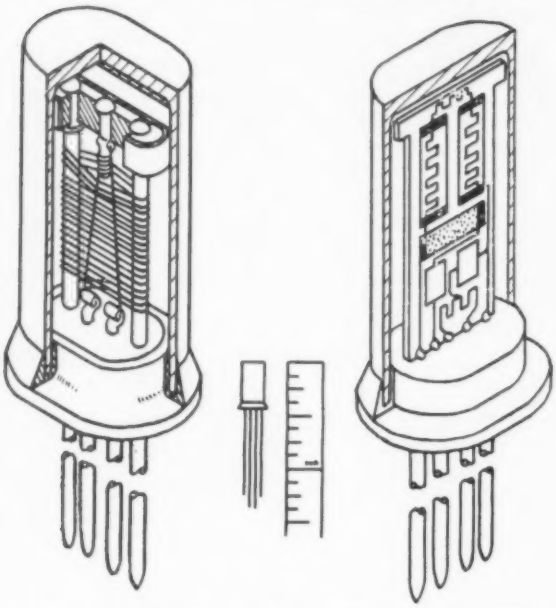


FIG. 9. Subminiature tube and component assemblies.

size of these tubes is approximately $0{\cdot}5 \times 0{\cdot}2$ in. diameter. Complete circuits can be made up of alternate component assemblies and tube assemblies, which take up very little space.

RELIABILITY

This field is beyond question the most important in the whole of electronics. Although it began to be recognized during the Pacific war, the difficulties in measurement, definition, etc., and the long time scale in any obvious remedial actions, have meant that only in the last few years has any effective action been taken. Some figures have been kept at R.R.E. on the failures of components in radar equipments undergoing environmental testing since 1944, and these are shown in Table 1. These figures can only be regarded as approximate as they are on small quantities of equipments. Larger numbers of equipments have been maintained at the Atomic Energy Research Establishment at Harwell, and an analysis of faults occurring

Table 1. Failures of components in radar equipments under environmental tests at R.R.E.

	Humidity tests only		Tests to Joint Service Spec. K114, Issue 1						Tests to Joint Service Spec. K114, Issue 2							
	1944	1945	1946	1947	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958
Total components	2983	4759	1617	1472	1871	1165	3150	1984	693	1456	1644	2393	5033	3673	2620
Total failures	106	137	58	67	40	28	66	19	16	27	25	29	43	71	26
Overall percentage failure*	3·6	2·9	3·6	4·5	2·1	2·4	2·1	1·0	2·3	1·9	1·5	1·2	0·9	1·9	1·0

* Average 2·1 per cent per approx. 1000 hr.

Table 2. Component failures under typical laboratory conditions at A.E.R.E., Harwell

	Percentage failure per annum*							
	1949	1951	1952	1953	1954	1955	1956	1957
Valves	4·2	3·3	4·7	4·0	3·2	2·5	4·96	6·1
Rectifiers	6·8	6·5	9·4	8·0	9·1	6·69	6·91	7·58
Double diodes and double triodes	3·5	3·0	4·0	4·2	4·1	4·8	5·42	7·2
Pentodes	5·8	2·9	4·2	2·6	2·67	5·6	5·96	7·99
Stabilizers	1·7	3·0	4·5	4·0	4·18	1·53	1·56	1·74
Miscellaneous	—	—	4·0	2·0	—	—	—	—
Resistors	0·54	0·27	0·20	0·30	0·39	0·25	0·2	0·21
High stability	0·45	0·37	0·30	0·4	0·73	0·64	0·45	0·43
Carbon composition	0·65	0·26	0·15	0·24	0·25	0·1	0·1	0·09
Wirewound	0·55	0·16	0·12	0·14	0·3	0·2	0·19	0·17
Potentiometers	0·11	0·26	0·27	0·24	0·22	0·19	0·19	0·19
Capacitors	1·1	0·19	0·31	0·14	0·16	0·13	0·33	0·36
Paper	1·9	0·20	0·37	0·14	0·22	0·15	0·15	0·11
Ceramic	0·45	0·06	0·08	0·14	0·14	0·04	0·04	0·08
Mica	0·18	0·04	0·1	0·08	0·12	0·09	0·03	0·1
Electrolytic	1·0	0·7	0·95	0·3	0·95	1·11	1·12	1·17
Transformers	18·0	2·6	1·8	1·4	2·0	0·9	0·99	0·83
Relays	1·9	0·48	0·68	0·54	0·52	0·3	0·4	0·62
Switches	0·88	0·31	0·45	0·32	0·22	0·2	0·22	0·19
Chokes	1·1	0·17	0·23	—	0·16	0·12	0·12	0·17
Metal Rectifiers	—	1·3	1·4	1·0	0·61	0·68	0·7	0·35
Meters	3·1	1·5	0·85	1·2	1·28	0·71	0·75	0·65
Totals (all components)	1·20	0·59	0·72	0·67	0·65	0·58	0·96	1·05

* Average 0·8 per cent (per approx. 1000 hr).

Table 3. Failures on temperature-stabilized, air-conditioned data-handling equipment at R.R.E.

	Diodes	Transistors	Capacitors	Resistors
Total no. of components	65,000	20,000	25,000	100,000
Time in use (hr)	1430	1400	1400	1400
Total no. of failures	26	59	1	3
Failure rate (%/1000 hr)	0·028	0·21	0·003	0·002

* Average 0·06 per cent per 1000 hr.

under typical laboratory conditions is given in Table 2.

Some recent figures on data-handling equipment at R.R.E. are shown in Table 3.

Development of reliable tubes, transistors and components

In the United Kingdom, development of a range of sub-miniature reliable tubes, wired directly into equipments, was commenced about 1952, primarily for use in guided missiles.

Considerable improvements have now been achieved in the failure rates previously encountered, shown in the following results compared with the normal commercial types per 1000 hr of operation:

Type	Electrical faults (%)	Mechanical faults (%)
Reliable range of tubes	0·3	0·3
Normal commercial range of tubes	1·0	3·0

These reliable tubes are purchased to a procurement specification, the limits of which are being progressively tightened as manufacturing control data is accumulated and analysed.

This is being followed by a range of components and also transistors. It is the intention to improve the reliability of components over a useful and workable range rather than on one or two particular capacitors and resistors; in order to be of use to an equipment designer it is essential to develop a range of reliable components which can fulfil 80 or 90 per cent of his requirements. This work was described by Mr. GRIFFIN at the 1959 Symposium, and has now reached the stage where some thirty-six different types of components are going through the five stages summarized below:

Stage 1. Of great importance is the agreement of a list of suitable components common to the missile industry.

Stage 2. Analysis of all available test data on manufacturers' components forming the common list.

Stage 3. Assessment of the selected products to determine in what manner further development is necessary.

Stage 4. Development contract to raise the standard of the product. Statistical estimate of the reliability of the components.

Table 4

Component	No. of styles
B.N.C. (R.F. high-temperature sealed) connector	20
Miniature rectangular multi-pole connector	6
Sub-miniature rectangular multi-pole connectors	6
Unitors	5
High-temperature multi-pole circular-sealed connector	8
Miniature multi-pole circular-sealed connector	8
Micro-miniature fully sealed connector	5
Micro-miniature relay, S/C. sealed 100*g*	1
Sub-miniature relay, 2×4 changeover, 100*g*	2
Sub-miniature relay, 6 changeover, sealed, 100*g*	1
As above but with latching facility	1
Sub-miniature relay, 8×12 changeover, sealed, 100*g*	2
Metallized paper capacitors, 24 values	12
Paper dielectric capacitors, 23 values	15
Mica dielectric capacitors, 27 values	2
Ceramic tubular capacitors, 7 values	2
Hi-K tubular capacitors, 3 values	2
Tantalum pellet-type capacitors, 27 values	4
Tantalum foil capacitors, 34 values	6
Trimmer capacitors, 4 values	2
Feed-thru capacitors, 8 values	1
Ledex switch	2
Wirewound resistors, 77 values	1
Precision wirewound resistors, 32 values	1
Grade 1 non-insulated resistors, 144 values	3
Grade 1 insulated resistors, 120 values	3
Grade 2 insulated resistors, 145 values	2
High-stability oxide-film resistors, 106 values	3
Wirewound potentiometer, 11 values	2
Pre-set phenolic-moulded potentiometer, 10 values	1
Trimmer potentiometer, 10 values	2

Stage 5. Procurement specification to very tight limits imposing quality and process control of manufacturing materials and processes. End-product non-destructive testing to eliminate potential early failures.

The types and ranges of components selected to date are shown in Table 4.

Acceptance quality levels are being discussed with each of the manufacturers concerned and a specification written to control production. So far nine specifications have been completed and it is hoped that the whole scheme will be completed within one year.

Whilst these components are being developed primarily for Guided Missiles, other users are expressing considerable interest and the prestige value of this "reliable" range is becoming of increasing importance.

Reliability evaluation of fixed resistors

It may be remembered that in a paper presented before the 1956 Symposium, an automatic machine for evaluating fixed-resistor reliability was described. This machine has now tested some 8000 resistors to the Joint Service Specifications RCS.112 and 114, and the results are summarized in Table 5.

The superiority of the oxide type of film resistor is thus established in the United Kingdom.

Some interesting work is being done in the development of oxide-coated glass-fibre resistors in which a 0·001 in. diameter glass fibre is coated with tin/antimony oxide film. A machine has been developed for the continuous coating of the fibre. Results to date show that resistance values of 10,000 Ω to 1 MΩ per inch of fibre are obtainable with temperature coefficients of better than 50 p.p.m./°C.

Miniature relays

Three electro-magnetic relays have recently been developed, which are shown in Fig. 10 (a)–(c).

A single-coil, single-changeover micro-relay, 0·3 in. diameter and 0·47 in. long, has been developed with the contact in a separate sealed compartment. Tests to date have shown a mechanical life exceeding 40,000,000 operations. At 15 W it has a life of 1,000,000 operations. Coils are available at 6, 9, 15, 30 and 50 V d.c., (the U.K. standard transistor component working voltages)

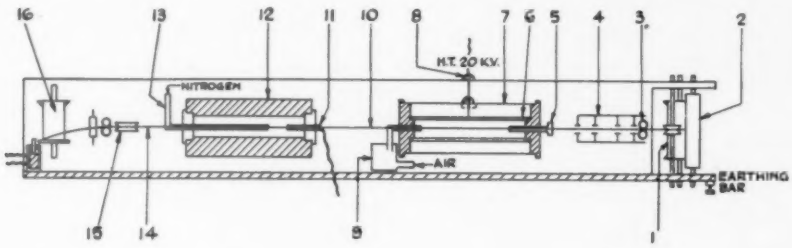
Table 5. *Summary of results of mass resistor testing at R.R.E.* (*each batch of* 1000 *from different manufacturers*)

Quantity tested	Type	Value (W)	Value (Ω)	Spec. limit (%)	Failures: Open circuit	Failures: Outside tolerance
1000	Carbon composition	¼	1500	10	0	4
1000	Carbon composition	¼	470,000	15	0	1
1000	Carbon composition	¼	150,000	15	0	1
1000	Cracked carbon	1	150,000	1·5	1	0
1000	Cracked carbon	1	100,000	1·5	4	31
1000	Cracked carbon	1	150,000	1·5	11	8
1000	Cracked carbon	1	150,000	1·5	28	50
250	Oxide film	1·4	1000	0·5	0	0
250	Oxide film		10 k	0·5	0	0
250	Oxide film		100 k	0·5	0	2
250	Oxide film		200 k	2·0	0	0
250	Oxide film	1·4	1000 k	0·5	0	0
250	Oxide film		10 k	0·5	0	0
250	Oxide film		100 k	0·5	0	0
250	Oxide film		200 k	0·5	0	0

at 300 mW. This cylindrical shaped relay consists of a pot-type magnetic circuit with armature "lid" to which is fixed the moving contact. The relay has withstood a level of more than 100 *g* vibration over a frequency of 50–2500 c/s. Operate and release times are less than 2 mμ, and bounce is below 1 μsec.

Two other relays [see Fig. 10(b) and (c)] have been designed specifically for missile use, using versing the coil connexions these become 14 and 20 msec. The two-coil relay is a postage-stamp type and measures $0{\cdot}9 \times 0{\cdot}8 \times 0{\cdot}4$ in.

A sub-miniature twelve-contact relay of minimum size and weight to withstand high-vibration conditions is under development and is shown in Fig. 10(d). A prototype indicates the feasibility of accomplishing this with a diameter of $\frac{1}{2}$ in., overall length of $\frac{3}{4}$ in., and a weight of 15 g. The twelve



1. SUPPLY BOBBIN WITH UNTREATED WIRE (EARTHED)
2. TENSIONING DEVICE WITH TRAVELLING PULLEY
3. DIRECTION DOLLIES
4. WIRE CLEANING BOX
5. ADJUSTABLE MASKS (THICKNESS CONTROL)
6. IONISING CAGE
7. PRECIPITATION CHAMBER (GLASS DUST FALLS THROUGH IONISING CAGE).
8. H.T. LEAD AND INSULATORS LEADING TO IONISATION CAGE.
9. DUST PROPELLING BOX (GLASS AND AIR)
10. WIRE COATED WITH UNFIRED GLASS DUST
11. THERMOCOUPLE
12. FURNACE (TO 1,000°C)
13. NITROGEN FEED
14. COATED WIRE
15. LOCATION FINDERS
16. POWER DRIVEN COLLECTING SPOOL (GEARED)

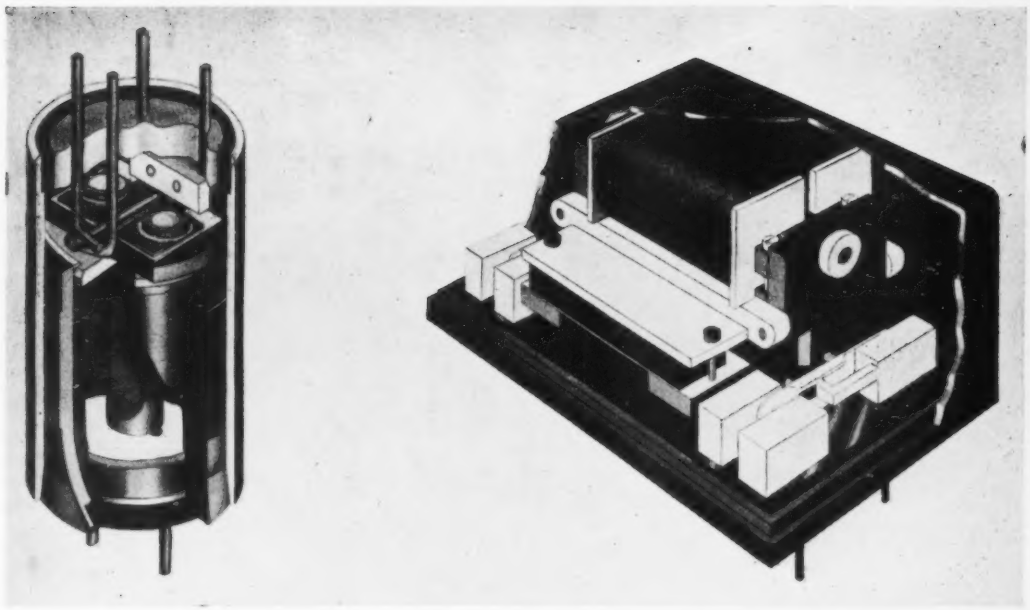
FIG. 11. Sketch of apparatus for coating wire with glass.

massed balanced rotating armatures linked together in such a way that neither can move without reciprocal and opposite movement of the other. The single-coil, single-changeover relay is $1{\cdot}1 \times 0{\cdot}726 \times 0{\cdot}7$ in., and instead of an armature return spring a permanent magnet is used to hold the armature in the operative position. The restrained springs of the spring-set arrangement assist in returning the armature when energization is removed from the coil. The spring sets then give a reasonably constant operate and release characteristic no matter what speed of energization is used. An interesting by-product of rotating armatures with permanent magnets is that reversal of polarity of source potential appreciably changes the operate and release times of the relay. The normal operate and release times for this relay are 5 and 10 msec respectively, but by reversing the coil connexions these become 14 and 20 msec. The two-coil relay is a postage-stamp type and measures $0{\cdot}9 \times 0{\cdot}8 \times 0{\cdot}4$ in.

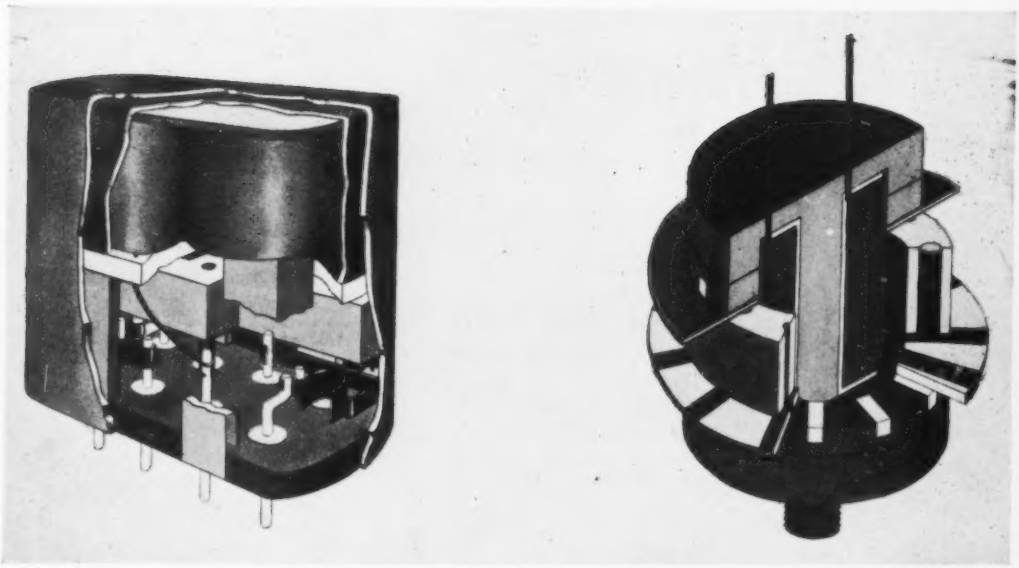
A sub-miniature twelve-contact relay of minimum size and weight to withstand high-vibration conditions is under development and is shown in Fig. 10(d). A prototype indicates the feasibility of accomplishing this with a diameter of $\frac{1}{2}$ in., overall length of $\frac{3}{4}$ in., and a weight of 15 g. The twelve contacts, consisting of suitably plated permanent magnets rated at $\frac{1}{2}$ A, 30V d.c., are contained within a hermetically sealed honeycomb. Contact gaps of the order of 0·005 in. at pressures of 5 g are considered sufficient to achieve a satisfactory life of a million operations.

Electrostatic precipitation of glass powders

Mention should also be made of a process under development at R.R.E. for coating metals with glass by means of electrostatic precipitation of a glass powder, the resultant deposit being fired to give a uniform coating, shown in Fig. 11. This has been used with some success to give wires a flexible thin coating for high-temperature insulation. The glass is matched in expansion coefficient to the chosen metallic substrate, to avoid crazing, etc. In this process the powder will penetrate into



(a) (b)



(c) (d)

FIG. 10. Recent designs of subminiature relays. (a) Single-changeover two-compartment micro-relay 0·3 in. dia., 0·47 in. long. (b) 100 *g* two-changeover relay 1·1×0·7 ×0·7 in. (c) Two-compartment, double-changeover relay 0·9×0·8×0·4 in. (d) Experimental magnetic slug relay, twelve-changeover, 0·56 in. dia., 0·6 in. long.

[*facing p.* 30

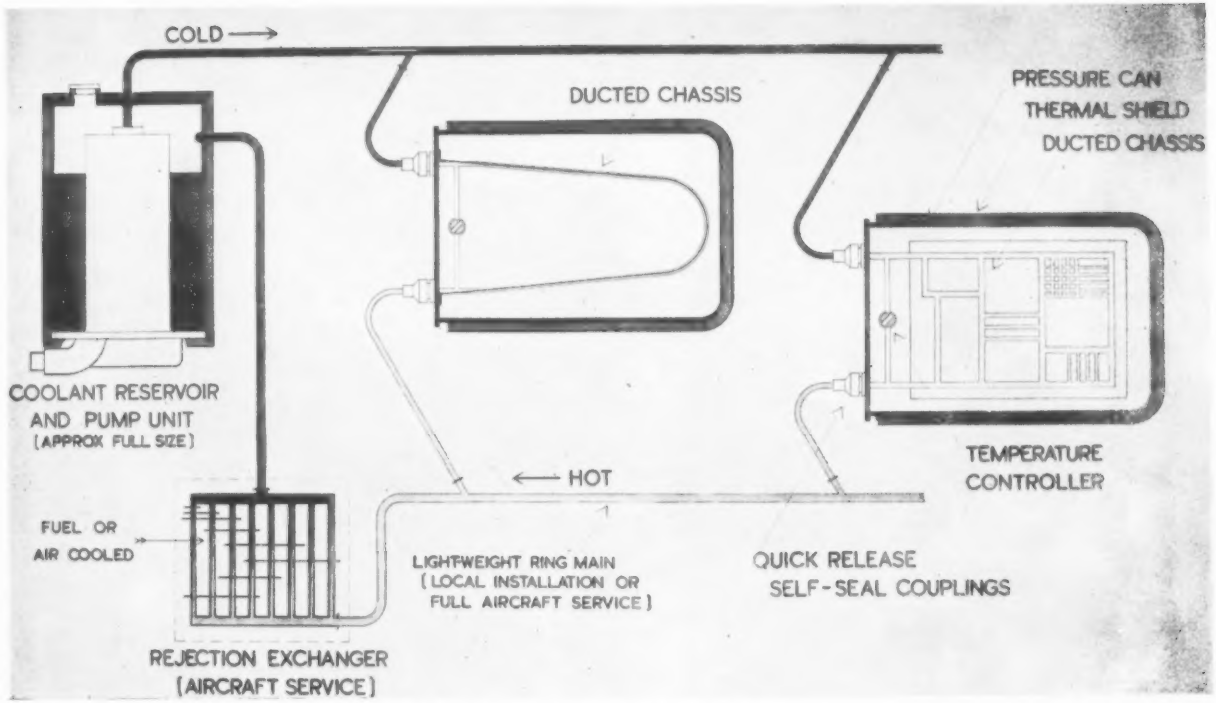


FIG. 12. Liquid cooling—schematic diagram.

crevices and can therefore be used to coat metallic pieces of irregular shapes.

Evaluation of hermetic-seal performance

Over 1000 seals (mainly glass/metal types) are being tested under d.c. and humidity conditions to reveal electromechanical corrosion effects, etc. Test voltages are −200, −100, −50, −5, +5, +50, +100 and +200V, and duration of humidity exposure from 7 to 84 days. Seals are considered to have failed when the insulation resistance falls to 100 Ω, and some interesting results are being obtained; for example, at the end of one series of 42-day tests, 73 out of 342 seals had failed. The results of these tests are being analysed and should be issued in a few months' time.

Deterioration of electronic components in storage

The problem of deterioration of components and materials in storage is a particularly acute one, and late in 1958 arrangements were made to store omponents over a period of 5 years. On the theory that reduction in temperature will delay chemical deterioration, three different temperatures were chosen for storage, +20°C, 0°C and −20°C. Both packaged and unpackaged items are included. The items being stored consist of 1440 fixed resistors, 310 variable resistors, 1340 fixed capacitors, 300 plugs and sockets, 160 relays, 100 transistors, 10 silicon rectifiers, 30 transformers and 40 potted assemblies, printed boards and wrapped joint assemblies.

Accurate measurements have been made on the main characteristics of all the components concerned and history cards are being kept of each type of component.

COOLING TECHNIQUES

Short-life rating of components

The importance of cooling on reliability is now well established. Tubes, transformers, resistors, capacitors, etc., all have decreased lives at high ambient temperatures, and results of some life/temperature investigations of the performance of certain resistors and capacitors at temperatures in the region at which they fail within a short time are briefly summarized.

Spiralled cracked carbon resistors remained for at least 100 hr within 5 per cent of their initial value at 200°C with either a.c. or d.c. loading. Some values and wattages remained within this figure at 250°C. Solid carbon composition resistors remained within 10 per cent of their initial values for 5 hr when fully loaded only at temperatures up to 150°C.

The insulation resistance of paper–foil and resin-impregnated metallized-paper dielectric capacitors falls rapidly with increasing temperature. If the initial insulation resistance of metallized-paper dielectric capacitors can be tolerated, then lives in excess of 10 hr can be expected at 150°C with full rated voltages applied. Failure is defined as a 10 per cent decrease in capacity, the insulation resistance increasing under these conditions. The capacitors can be operated at a maximum temperature of 170°C if less than 10-hr life is acceptable. The effect of a reduction in applied voltage is to decrease the insulation resistance during the first part of life. It later recovers and continues to increase. The rate of decrease in capacitance is less than with full applied voltage. The maximum operating temperature for foil–paper capacitors is in the region of 150°C when a life in excess of 10 hr will be obtained before a sudden short or open circuit, the capacitance still remaining within ±10 per cent of its initial value. Capacitors with the lowest initial insulation resistances generally had the shortest lives.

This work was done for the missile industry in order to give some idea of lives to be expected at temperatures at higher than normal, and also to give an indication of the mechanism of breakdown at high temperatures.

Cooling of equipments

The cooling of electronic equipment is a complex subject. Extreme temperatures, especially high temperatures, undoubtedly lead to unreliability and it therefore seems logical to aim for a system of temperature stabilization if maximum reliability is to be realized.

After a study of the many methods of cooling, it is considered that there is a strong case for developing the indirect liquid-cooling system as a means of temperature stabilization, particularly for military airborne electronic equipment. The system should be satisfactory over the ambient temperature range −60°C to +150°C at least, and has development potential for the application of evaporative methods.

It is proposed to insulate the equipment completely from its surroundings, and then circulate the coolant either through the chassis or through the walls of the cover, or through both. Ducted aluminium is an excellent material for chassis and covers, and R.R.E. was probably the first to apply it for this purpose in the United Kingdom. The coolant is a silicate ester based material, Montosil "E", which is similar to the U.S. "Coolanol 45"; it is preferred to water–glycol mixture for several reasons, e.g. freedom from corrosion, wide temperature range, less "searching" on joints and in the circulating system. The ducted aluminium is therefore the collection area for component waste heat, which will be transferred mainly by conduction and convection modes. It is not possible to get cooling closer to the components than the chassis itself and the liquid acts as the transporter, taking the heat energy away economically from the electronic units to a remote heat exchanger, which is provided as an aircraft service. The flow rates and pressures are reasonable, about 3 cm/W per min at 15 lb/in^2 delivery from a submerged pump consuming less than 5 per cent of the heat energy removed. A schematic diagram of the system is shown in Fig. 12.

It is proposed to control the temperature of equipments individually by regulating coolant flow rates on the outflow temperature, equipments being brought to stabilized operating temperature quickly, because of the external insulation. Quick-release, self-sealing couplings are used for the coolant so that there are no problems of draining the equipment before disconnecting. It is as easy as disconnecting a plug and socket. For maintenance and tuning, the equipment is connected to a single-liquid circulating system in the workshop. Peak performance is guaranteed because circuits and components operate at the same ideal temperature as they will in the aircraft, irrespective of the aircraft ambient temperature. It is fully appreciated that there are other cooling systems which have particular merits, but it is felt that this indirect liquid-cooling system with its insulated containers and temperature-stabilized electronic circuits may prove to be a simple reliable method for many installations. Controlled life tests are being conducted on two sets of airborne radar equipment, one fitted with liquid-cooled covers, the other with normal covers and cooled by natural convection. Aircraft environmental conditions are simulated and it is intended to run the equipments until it is no longer economic to repair them—certainly several thousand hours. This experiment should give interesting data on components and circuit reliability, and indicate the effect of temperature stabilization on equipment reliability.

Cooling of microminiature assemblies

Perhaps, surprisingly for circuits with such small power consumption, the actual heat densities in microminiature units are quite high, and on thin-glass or ceramic substrates local hot spots are possible. It may be wise to control the temperature of the complete microminiature equipment (i.e. one sealed container containing a number of stacks of wafers). Again, ducted aluminium may be used for a completely insulated cover, with a fluid circulating through it.

Miniature wrapped joints

Inside equipment containing microminiature assemblies it is proposed to fit a number of sub-units, each containing a stack of wafers. A technique for assembling the sub-units using a very thin miniature wrapped joint is being developed.

In the United Kingdom most encouraging results have been obtained with wrapped joints as reliable connexions, in particular in an application to a supersonic aircraft radar. Wrapped joints, using terminals of monel metal, have been used instead of plugs and sockets, also for connexion of components operating at audio and video frequencies. These wrapped joints are also being used for interconnexion of sub-units in a missile system. The Bell Telephone Laboratories' claims for wrapped-joint reliability are supported by our experience in military equipments and a very small version (known at R.R.E. as "Miniwraps") is being developed for this application. The initial target is 100 joints to the square inch, using 36–40 British gauge wrap wire on terminals approximately 10 by 25 mils. A cored copper is preferred so that a relatively high wrapping tension can be used. The following advantages are anticipated:

(1) Completely uniform cold-formed connexions.
(2) Potential reliability better than for soldered or alloyed connexions because of the re-

duced number of variables in the jointing process.

(3) Suitable for automatic machine assembly.

(4) Facility to change a faulty wafer.

The arms of a U-link of wire are simultaneously wrapped on to the two terminals to be connected, using a twin-headed wrap tool. An alternative is to bind together two rectangular terminals, i.e. a bound joint. This type of connexion has been used satisfactorily in ground telecommunications equipment in the U.K. to bind round component leads against rectangular terminals, and it has been shown by severe environmental tests and radioactive tracer tests that these bound joints are reliable—more so than might be expected on geometrical considerations. It is indeed yet another pressure joint with a store of energy to maintain the good connexion.

A plastic moulding is proposed for the sub-unit, with end slots at 0·1 in. centres to locate the wafers, as shown in Fig. 13. The top of the sub-unit is made of aluminium with a thin silastomer coating, which will be locally compressed by the glass wafers; this should cushion them and provide good thermal contact. Wafer dimensions have not yet been standardized, but for an experimental

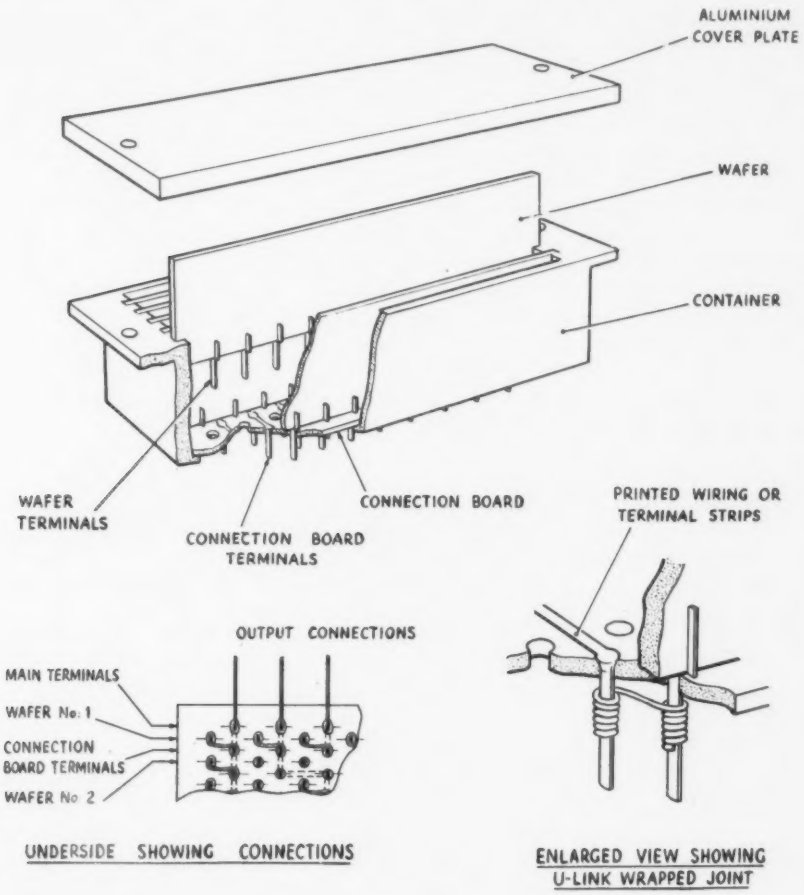


FIG. 13. Proposed wrapped-joint assembly system.

project the wafer is 2·6 × 0·6 in. with a circuit area 2·5 × 0·5 in. and ten terminations. The heat-conduction distance from components is thus not more than about $\frac{1}{4}$ in. in a 0·040 in. material. This is better for cooling surfaces than a square wafer of equivalent surface area. If natural convection between the wafers maintained by the cold aluminium top surface (plus conduction) is insufficient to prevent hot spots, air or inert gas can be circulated between the wafers for forced convection cooling.

An alternative scheme for using circuits on 0·5 in. square plates is to enclose them in thin metal containers, sealing them to bottom plates by half-cycle a.c. heating which confines the heat to a small area. The square-section terminals project through ceramic–metal seals, and would be connected by wrap joints similar to Fig. 3. In this case all units would be approximately 0·6 × 0·6 × 0·2 in. and possibly filled with silicone grease to aid heat dissipation.

The miniature wrapped joint proposal is particularly interesting as a means of automatically assembling micro-miniature wafers. The terminals are located on a 0·1 in. module and it should therefore be practical to tape instruct the wrapping tool to proceed to each position in sequence for making connexions. Similarly, the removal of a faulty wafer involves the unwrapping of one row of connexions without disturbing the other wafers and connexions of the sub-unit. It is envisaged that sub-units can also be interconnected by wrapped-joint links, so that if necessary a complete sub-unit could be replaced.

CONCLUSION

This paper has necessarily covered a wide field, but it is hoped that this review of U.K. work will be of interest. The author would like to acknowledge the work of his colleagues at R.R.E., and of many others, in its preparation.

REFERENCE

1. R. H. Alderson and F. Ashworth, *Brit J. Appl. Phys.* **8**, 205 (1957).

Solid-State Electronics Pergamon Press 1961. Vol. 2, No. 1, pp. 35–67. Printed in Great Britain

PROBLEMS RELATED TO p–n JUNCTIONS IN SILICON†‡

WILLIAM SHOCKLEY

Shockley Transistor, Unit of Clevite Transistor, Palo Alto, California

Abstract—The phenomena of secondary ionization, avalanche breakdown and microplasma phenomena in p–n junctions are analyzed using a simplified model in which holes and electrons have identical properties described by four constants. Only two scattering processes for carriers are considered, each having two constants: generation of highest energy or Raman phonons, energy E_R and mean-free-path L_R; ionization or electron–hole pair production, threshold carrier energy E_i and mean-free-path L_i. E_R is determined from neutron scattering data; $E_R = 0{\cdot}063$ eV for Si and $0{\cdot}037$ eV for Ge. The other three constants are adjustable. E_i and $L_i/L_R = r$ are chosen to fit data on quantum yield for photons with $1 < h\nu < 5$ eV using the formula $Q = 3-2\exp(E_g+2E_i-h\nu)/2rE_R$. For silicon this gives $E_i = 1{\cdot}1$ eV (which is equal to the energy gap E_g) and $r = 17{\cdot}5$. For germanium E_i is also about $1{\cdot}1$ eV and $r = 57$. The simple model predicts that the ionization coefficient $\alpha(F)$ varies with field F as $(qF/rE_R)\exp -(E_i/qL_RF)$ which is in good agreement with CHYNOWETH's data for electrons in silicon if L_R is set equal to 50 Å. The model predicts an energy per pair for ionization by high-energy particles of about $2{\cdot}2\ E_i+rE_R$ which is in good agreement with measured values. It also predicts a hot-carrier random energy of about $0{\cdot}2$ eV for $F = 400{,}000$ V/cm, which agrees with the spectra of hole–electron recombination in microplasmas. Thus the three adjustable constants permit fitting six pieces of experimental data in four independent experiments in spite of the fact that the intricacies of the band structure are disregarded.

The effects of statistical spatial fluctuations of donor and acceptor ions are considered and it is concluded that these will be randomly distributed according to a Poisson distribution. This randomness leads to a characteristic fluctuation voltage $(qF_B/K)^{1/2} \simeq 0{\cdot}3$ V for silicon where F_B is the breakdown field, and the dielectric constant $K = 1{\cdot}04\times10^{-12}$ F/cm for silicon. The effect of these fluctuations is to produce local regions in a p–n junction with breakdown about $0{\cdot}7$ V lower than the average in uncompensated material. The fluctuations of voltage are larger by $[(N_d+N_a)/(N_d-N_a)]^{1/2}$ in compensated material. The fluctuations can increase the apparent ionization coefficient substantially.

Microplasma effects are considered and it is shown that in a junction with only the Poisson fluctuations the microplasma should be stabilized by an apparent series resistance due to space charge of magnitude $1/v_{max}K \simeq 10^5\ \Omega$ where $v_{max} = (E_R/m^*)^{1/2}$ is the limiting drift velocity. This is much larger than the spreading resistance term of magnitude $1/\mu F_BK \simeq 2000\ \Omega$. It is concluded that typical noisy microplasma phenomena are probably associated with localized structural defects probably having two characteristics: (1) they increase the effect ionization coefficient to a value greater than 10^5 cm^{-1} over a region less than 10^{-5} cm long; (2) they have a mechanism for capturing charge which increases the field once the microplasma has formed. Small SiO_2 precipitates and dense arrays of dislocations appear to have the requisite properties.

Metal precipitates in the space-charge layer produce "soft" reverse characteristics with localized currents of the form $V^{6\pm1}$.

Résumé—Un modèle simplifié, où électrons et trous ont des propriétés identiques décrites par quatre constantes a été employé pour analyser les phénomènes d'ionisation secondaire, de microplasma et de rupture à avalanche. Seuls deux procédés de diffusion de porteurs, ayant chacun deux constantes, sont considérés: génération de phonons Raman ayant la plus grande énergie (énergie E_R et libre parcours moyen L_R); ionisation ou production d'électrons-trous (limite d'énergie de porteur E_i et libre parcours moyen L_i). E_R est déterminée par les données de diffusion des neutrons; $E_R = 0{,}063$ eV pour le Si et $0{,}037$ pour le Ge. Les trois autres constantes peuvent être ajustées. E_i et $L_i/L_R = r$ ont été choisies de facon à s'adapter au rendement quantique des photons d'après la formule $Q = 3-2\exp(E_g+2E_i-h)/2rE$, avec la limite $1 < h < 5$ eV. Ceci donne pour le silicium,

† The major part of the research reported here was carried out under a contract with the Air Force Cambridge Research Laboratories, U.S.A.

‡ Address delivered at International Conference on Semiconductor Physics, Prague, Czechoslovakia, August 1960.

E_i = 1,1 eV (valeur égale a l'intervalle d'énergie E_g) et r = 17,5. Pour le germanium, E_i est aussi environ 1,1 eV et r = 57. Le modèle simplifié prédit que le coefficient d'ionisation $\alpha(F)$ varie en fonction du champ F d'après l'expression $(qF/rE_R)\exp -(E_i/qL_RF)$, ce qui est en accord avec les données de CHYNOWETH pour les électrons dans le Silicium si L = 50 Å. Le modèle prédit une valeur de $2{,}2E_i+E_R$ pour l'énergie de chaque paire obtenue par ionisation de particule à haute ènergie; ce résultat s'accorde avec les valeurs mesurées. Il prèdit aussi une énergie aléatoire de porteur à chaud d'environ 0,2 eV pour F = 400.000 V/cm, en accord avec le spectre de recombinaisons trou-electron des microplasma. Ainsi les trois constantes qui peuvent être choisies permettent d'ajuster six données expérimentales dans quatre expériences indépendantes malgrè que l'on ignore les complications de la structure de bande.

Les effets des fluctuations statistiques d'espace des ions donneurs et accepteurs sont analysés et on conclut que ceux-ci seront distribués aléatoirement en accord avec une distribution de Poisson. Cette tendance aléatoire menè a une tension caractéristique de fluctuation d'après l'équation $(qF_B/K)^{1/2} \simeq 0{,}3$ V pour le Silicium, F_B étant le champ de rupture et K la constante diélectrique $(+1{,}04\times10^{-12}\ F/\text{cm})$. L'effet de ces fluctuations est de produire des régions localisées dans une jonction p–n ayant des tensions de rupture d'environ 0,7 V de moins que la moyenne dans un materiau non-compensé. Les fluctuations peuvent augmenter sensiblement le coefficient d'ionisation apparent dans un matériau compensé, le facteur d'augmentation des fluctuations étant $[(N_d+N_a)/(N_d-N_a)]^{1/2}$ dans ces mêmes conditions.

Les effets microplasma sont considérés et on démontre que dans une jonction, ne comprenant que les seules fluctuations de Poisson, le microplasma doit être stabilisé par une resistance en série apparente dûe à la charge d'espace de valeur $1/v_{max}$. $K \simeq 10^5\ \Omega$ ou $v_{max} = (E_R/m^*)^{1/2}$ est la limite de la vitesse d'apport. La resistance d'extension donnée par la formule $1/\mu F_B K \simeq 2000\ \Omega$ est beaucoup plus petite. On conclut que les phénomènes des microplasmas bruyants typiques sont peut-être associés avec des défauts de structure locaux ayant probablement les deux caractéristiques suivantes: (1) Ils augmentent le coefficient d'ionisation effectif d'une valeur plus grande que 10^{-1} cm sur une surface moins longue que 10^{-5} cm; (2) ils possèdent un mécanisme pour capturer des charges qui augmentent le champ, une fois le microplasma formé. De petits précipités de SiO_2 et des assemblages denses de dislocations semblent avoir les propriétés requises.

Des precipités de métal dans la couche de charge d'espace produisent des caractéristiques ramollies dans le sens inverti avec des courants localisés d'après $V^{\theta\pm1}$.

Zusammenfassung—Die Phänomene der sekundären Ionisation, des Lawinenzusammenbruchs und Mikroplasma-Erscheinungen in p–n–Übergängen werden an einem einfachen Modell erörtert, in dem Löcher und Elektronen identische, durch 4 Konstante bestimmte Eigenschaften besitzen. Nur zwei Streuprozesse für Träger werden berücksichtigt, jeder mit zwei Konstanten: die Erzeugung von Phononen höchster Energie (Raman–Phononen), Energie E_R, und die mittlere freie Weglänge L_R; die Ionisation oder Elektron-Loch-Paar Erzeugung, Schwellenenergie der Träger E_i, mittlere freie Weglänge L_i. E_R wird durch Neutronenstreuung bestimmt; E_R = 0,063 eV für Si und 0,037 eV für Ge. Die übrigen drei Konstanten sind regulierbar. E_i und $L_i/L_R = r$ werden so gewählt, dass sie der Quantenausbeute für Photonen ($1 < h < 5$ eV) entsprechen, wobei die Formel $Q = 3-2\exp(E_g+2E_i-h)/2rE_R$ benutzt wird. Für Silizium ist E_i = 1,1 eV (gleich der Energielücke E_g), r = 17,5. Für Germanium ist E_i ebenfalls etwa 1,1 eV und r = 57. Nach dem einfachen Modell soll der Ionisationskoeffizient $\alpha(F)$ mit den Feld nach folgender Beziehung variieren: $(qF/rE_R)\exp -(E_i/qL_RF)$. Dies steht in guter Übereinstimmung mit CHYNOWETHS Angaben für Elektronen in Silizium, wenn man L_R = 50 Å setzt. Die Energie eines Paares bei Ionisation durch hochenergetische Teilchen soll nach dem Modell etwa $2{,}2E_i+rE_R$ betragen, was mit den Messungen gut übereinstimmt. Die statistische Energie der heissen Träger sollte etwa 0,2 eV sein, wenn F = 400 000 V/cm. Dies stimmt mit den Spektren der Rekombination von Löchern und Elektronen im Mikroplasma überein. Mit Hilfe der drei regulierbaren Konstanten lassen sich also 6 experimentelle Messungsergebnisse in 4 unabhängige Versuche einfügen, obwohl man die komplexe Bandstruktur nicht berücksichtigt.

Die Effekte statistischer räumlicher Schwankungen der Donator und Akzeptorionen werden in Betracht gezogen, mit dem Ergebnis, dass diese nach einer Poissonschen Verteilung statistisch angeordnet sind. Dies führt zu einer charakteristischen Schwankungsspannung $(qF_B/K)^{1/2} \simeq 0{,}3$ V für Silizium, wo F_B das Feld beim Zusammenbruch und K die Dielektrizitätskonstante $+1{,}04\times$ $\times10^{-12}\ F$/cm für Si ist. Durch diese Schwankungen entstehen in einem p–n–Übergang lokale Zonen, deren Zusammenbruchsspannung um etwa 0,7 V niedriger ist als der Durchschnitt in nicht kompensiertem Material. Hier sind die Spannungsschwankungen um $[(N_d+N_a)/(N_d-N_a)]^{1/2}$ grösser. Diese Schwankungen können den scheinbaren Ionisationskoeffizienten beträchtlich erhöhen.

Mikroplasmaerscheinungen werden dann erörtert, und es wird gezeigt, dass in einem Übergang,

in dem nur Poisson-Schwankungen auftreten, das Mikroplasma durch einen scheinbaren Serienwiderstand stabilisiert werden sollte, der von einer Raumladung der Grösse $1/v_{max}K \simeq 10^5\ \Omega$ herrührt, wo $v_{max} = (E_R/m)^{1/2}$ der Grenzwert der Driftgeschwindigkeit ist. Dieser Wert ist viel grösser als der Ausbreitungswiderstand von $1/\mu F_B K \simeq 2000\ \Omega$. Typpsche Rauscherscheinungen im Mikroplasma werden lokalisierten Strukturdefekten zugeschrieben, die wahrscheinlich zwei Kennzeichen haben: (1) Sie erhöhen den effektiven Ionisationskoeffizienten auf einen Wert von über $10^5\ cm^{-1}$ über eine weniger als 10^{-5} cm lange Zone. (2) Sie sind imstande, Ladung einzufangen, die nach Bildung des Mikroplasmas das Feld verstärkt. Kleine Niederschläge von SiO_2 und dichte Anordnungen von Dislokationen scheinen die erforderlichen Eigenschaften zu besitzen. Metallniederschläge in der Raumladungsschicht erzeugen „weiche" Umkehrcharakteristiken mit lokalisierten Strömen der Form $V^{6\pm1}$.

I. SCIENCE AND TECHNOLOGY AND p–n JUNCTIONS

THE phenomena occurring in reverse-biased p–n junctions in silicon are of interest both scientifically and technologically.

The phenomena selected for discussion are concerned with the generation of hole–electron pairs by high-energy holes and electrons.

In Section II a simple model for this process is discussed and is shown to tie together a wide group of experimental observations. Some mysteries having to do with discrepancies between different sets of measurements are introduced.

In Section III the model for the reverse-biased p–n junction is examined from the point of view of the statistical fluctuations in electric field due to randomness in the arrangement of donors and acceptors. The effect of these fluctuations on junction properties is examined, and it is concluded that they may explain the mysteries raised in Section II.

In Section IV data on multiplication are examined, and it is concluded that many previous measurements of secondary multiplication have not been distorted by gross variations in junction properties from region to region.

Section V deals with the problem of microplasma formation. The statistical fluctuations discussed in Section III are analyzed and their predicted effects calculated. It is concluded that structural imperfections, such as SiO_2 precipitates, probably cause noisy microplasmas to form.

Section VI discusses the effects produced by metal precipitates.

Section VII states the conclusions.

II. A SIMPLE MODEL FOR SECONDARY IONIZATION EFFECTS IN SILICON

II A. *The empirical character of the model*

The model for secondary ionization discussed in this section should be regarded as phenomenological or empirical, rather than theoretical. A theoretical model should take into account the energy-band structure for energies greater than 1 eV from the band edges and should make use of the appropriate effective masses for hot-carrier effects. No attempt has been made here to introduce these refinements.

The model discussed here uses three adjustable constants. These are adjusted to fit data on quantum yield and secondary ionization by electrons. The model with the three adjustable constants determined in this way is then found to give reasonable agreement with two other independent experiments: the effective temperature of recombination radiation in microplasmas and the average energy per hole–electron pair formed by a high-energy ionizing particle or a photon.

Altogether the model with three adjustable constants can be made to fit six experimental values involved in four independent experiments. This agreement would appear to give the model a high degree of validation if it were based upon a sound theoretical foundation.

The fact that the model greatly oversimplifies the physical phenomena involved, however, suggests that the three adjustable constants do not have the simple physical significance that the model assigns to them, but are instead some sort of complex average of a variety of related effects.

It is to be hoped that a more complete theory for the interesting electronic phenomena in p–n junctions at high reverse bias will be developed and it can be seen how the quantities involved are related to the constants of the simple model.

II B. *The four-parameter model with three adjustable constants*

For carriers in fields so high that secondary ionization processes are significant it is probable

that the dominant scattering by phonons involves the phonons of highest frequency.[1,2]

The highest energy phonons are those in which the two face-centered cubic sub-lattices of the diamond structure are vibrating in opposite directions. The wave vector for this vibration has zero length and corresponds to the center of the Brillouin zone. The vibration is three-fold degenerate. This vibration has been referred to as the "Raman" vibration for the crystal. The energy of a quantum at this frequency is denoted in this work by E_R.

Values of E_R have been determined by neutron diffraction for silicon by PALEVSKY *et al.*[3] and for germanium by BROCKHOUSE and IYENGAR[4]. The values are

$$\begin{aligned} E_R &= 0{\cdot}063 \pm 0{\cdot}003 \text{ eV for Si} \\ &= 0{\cdot}037 \text{ eV for Ge} \end{aligned} \tag{2.1}$$

For non-vanishing vectors in the Brillouin zone, the energy decreases, falling to about $0{\cdot}7\ E_R$ at the edge of the zone. In the model treated here, the energy of the high-frequency phonons is taken as constant and equal to E_R. All collisions are considered to produce energy losses, which is a good approximation for Si at room temperature with $kT < E_R/2$ so that the high-frequency modes are not excited.

E_R is one of the four parameters of the model but since it is independently determined it is not an adjustable constant.

The three adjustable constants are:

E_i = threshold energy measured from the band edge for a carrier above which it may produce a hole–electron pair (2.2)

L_R = the mean-free-path between scattering by "Raman" modes (2.3)

$r \equiv L_i/L_R$ where L_i is the mean-free-path between ionizations for a carrier with energy greater than E_i. (2.4)

From definition (2.4) it follows that an electron with energy greater than E_i generates on the average r phonons per ionization.

The difference between the simple model and the previous model of secondary effects can be appreciated by considering the probability that a carrier introduced into the crystal with energy E can produce ionization before dissipating energy $E-E_i$ to phonons and becoming unable to ionize. In order to lose this energy it must make

$$C = (E-E_i)/E_R \tag{2.5}$$

collisions with phonons. It must thus move on the average a distance CL_R. The probability that it does not ionize in this distance is:

$$\begin{aligned} &\text{Probability of not ionizing while} \\ &\text{losing energy } E-E_i \text{ to phonons} \\ &\qquad = \exp(-CL_R/L_i) \\ &\qquad = \exp -(E-E_i)/rE_R \end{aligned} \tag{2.6}$$

The probability of producing one (or more) ionizations in slowing down from E to below E_i is thus:

$$\begin{aligned} P(E) &\equiv \text{Probability of ionization} \\ &= 1 - \exp -(E-E_i)/rE_R \end{aligned} \tag{2.7}$$

This probability is plotted in Fig. 1 using the values derived in subsection II C with $E_i = 1{\cdot}1$ eV and $rE_R = 1{\cdot}1$ eV.

Also shown in Fig. 1 is the previous theory. It represents a probability of ionization that is zero for energies below a threshold of about 2·3 eV and unity for higher values. This is the model used by WOLFF[2] in calculating the secondary ionization coefficient. It in turn represents a simplification, as WOLFF points out, for a situation in which the probability rises from zero toward unity over a small energy range. We shall return to the interpretation of the dependence of the probability of ionization upon energy in subsection II C.†

II C. *Application of the simple model to enhanced quantum yield for high-energy photons*

High-energy photons produce photo-electrons and photo-holes of sufficiently high energy that these in turn produce hole–electron pairs. Two ranges of photon energies are of particular interest: the very high-energy range in which many generations may be involved, and the low-energy range in which the photocarriers just begin to generate pairs and the quantum yield is enhanced

† *Note added in proof:* It is probable that the very short mean free path and high probability of ionization used by WOLFF does apply above some threshold energy. This energy may be higher than 2·3 eV. Thus a correct curve would be represented at each energy by the higher of the two theories of Fig. 1.

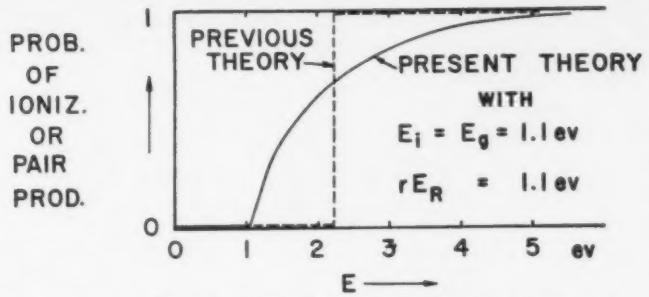


FIG. 1. Comparison between previous theory and present simple model of probability of ionization as a function of carrier energy.

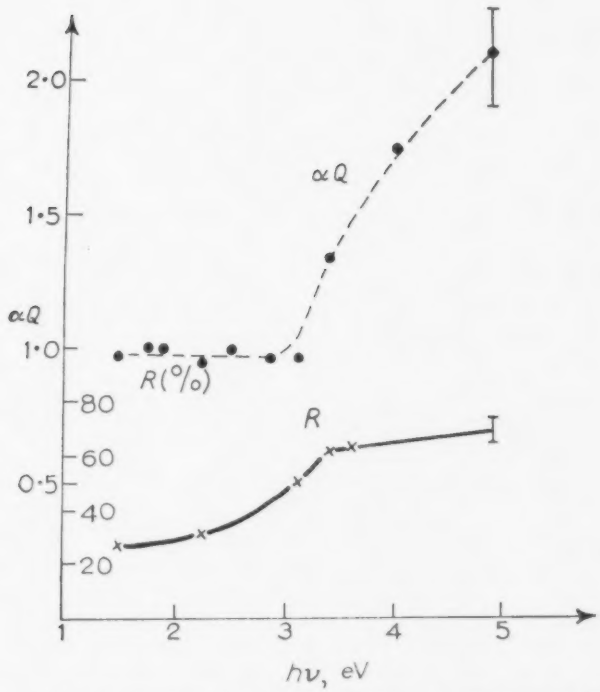


FIG. 2. Quantum yield, denoted by αQ, and reflection coefficient R in silicon as a function of photon energy (VAVILOV[5]).

above unity. This subsection is concerned with the latter effect. High-energy effects are discussed in subsection II E.

The threshold for enhanced quantum yield and the yield for a few electron volts above this threshold have been studied for silicon by VAVILOV[5], VAVILOV[5], TAUC[6] and KOC[7,8] have studied germanium, while TAUC and ABRAHAM[9] have measured indium antimonide. Figs. 2 and 3 show the data of VAVILOV.

The simple model can be used to fit the data of VAVILOV for silicon with the aid of the simplifying assumption that the excess energy $h\nu - E_g$ of the photon is divided equally to give the photo-hole

and photo-electron each an energy of $(h\nu - E_g)/2$. For the high energies of interest, it is assumed that direct transitions are involved so that no phonon energy need be subtracted from $h\nu$. Each photo-carrier may then ionize if this energy is greater than E_i. In accordance with the results of the preceding subsection each of the photo-carriers has a

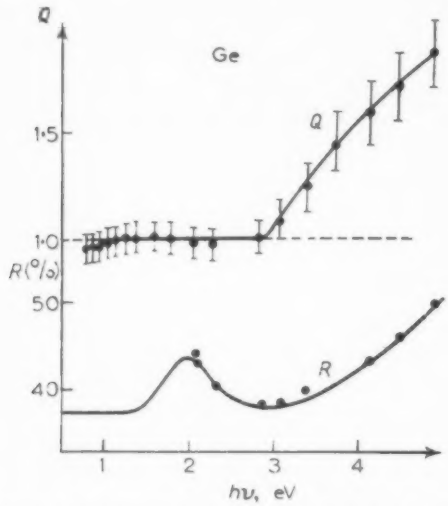


FIG. 3. Quantum yield in germanium as a function of photon energy (VAVILOV[5]).

probability $P(E)$ of producing one (or more) additional pairs where

$$P[(h\nu - E_g)/2] = 1 - \exp -(h\nu - E_g - 2E_i)/2rE_R \tag{2.8}$$

Over the range of $h\nu$ in which "third generation" pairs can be neglected, the quantum yield should thus be $Q(h\nu)$ where

$$\begin{aligned} Q(h\nu) &= 1 + 2P[(h\nu - E_i)/2] \\ &= 3 - 2\exp -(h\nu - E_g - 2E_i)/2rE_R \end{aligned} \tag{2.9}$$

This gives a threshold energy of

$$h\nu = E_g + 2E_i \tag{2.10}$$

for the photons.

Fig. 4 shows a fit of VAVILOV's data for silicon using $E_g = 1{\cdot}1$ eV as known and choosing the adjustable constants as:

$$E_i = 1{\cdot}1 \text{ eV and } rE_R = 1{\cdot}1 \text{ eV} \tag{2.11}$$

It is seen the fit is excellent. Using the value 0·063 eV for E_R leads to

$$r = 17{\cdot}5 \text{ phonons generated per ionization} \tag{2.12}$$

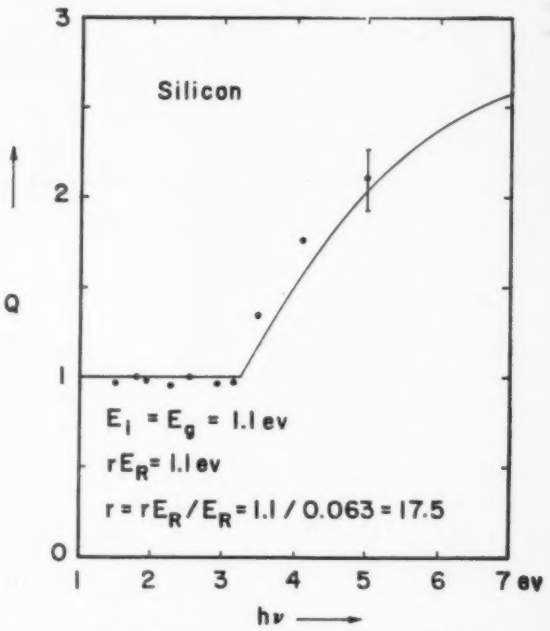


FIG. 4. A fit of the present simple model to the silicon data of VAVILOV[5].

Fig. 5 shows a similar fit for germanium taking $E_i = E_g = 0{\cdot}66$ eV. It is seen that the fit is not as good and that a larger value of E_i would give a better fit. The value of rE_R is 2·1 eV for $E_R = 0{\cdot}037$ eV, corresponding to $r = 57$.

There are several puzzling aspects about the data and the fit obtained by the simple model. In the first place, there is no *a priori* reason why the energy should split equally between the photo-hole and the photo-electron in silicon. On the other hand, both the silicon and the germanium data suggest that only one threshold is involved in the energy range studied. (For indium antimonide, on the other hand, two thresholds are clearly seen, TAUC and ABRAHAM[9].) The fact that the silicon data run smoothly to $Q > 2$, suggests that both the photo-hole and the photo-electron can produce pairs. The germanium data do not suggest a second

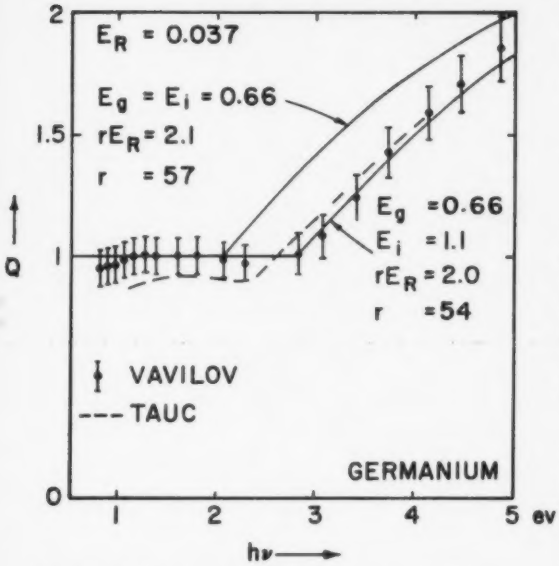


FIG. 5. A fit of the present simple model to the germanium data of VAVILOV[5] and of TAUC[6].

threshold, and although they do not extend beyond $Q > 2$ they fall on a curve which can be best fitted with a yield of the form

$$Q = 1+x[1-\exp(h\nu-h\nu_0)/2rE_R] \quad (2.13)$$

if x is chosen between 1·5 and 2·5. This suggests that both photocarriers are ionizing for $h\nu$ above 2·9 eV in germanium.

A theory of enhanced quantum efficiency has been given by ANTONČÍK[10,11]. This differs in an essential way from that presented here. According to ANTONČÍK's model, the variation quantum yield arises from the distribution of photo-electrons in the Brillouin zone immediately after photo-ionization. The variation in yield arises from the variation in the number which are in the part of the zone where they can produce impact ionization.

In this work ANTONČÍK considers, but rejects, the possibility that scattering of photo-electrons by phonons into the proper part of the zone may affect ionization probabilities. The simple model may have some validity if this rejected mechanism is actually an important one. The quantity r might then represent the ratio of volume where ionization was not possible to volume where it was possible.

Another possible interpretation of the large values for r is that all of the enhanced quantum yield arises from processes in which phonons are used to conserve momentum and hence are less probable than phonon generation.

(See also footnote p. 38.)

II D. *Application of the simple model to secondary multiplication*

A continuous improvement in the accuracy of the measurement of α, the secondary multiplication coefficient (defined as pairs produced per cm in a field of F V/cm) for holes or electrons, has occurred since the original work of MCKAY and MCAFEE[12]. The most recent, and probably most accurate, data are those of CHYNOWETH[13] for electrons in silicon. These data are shown in Fig. 6.

Two particular features of the data should be noted for the purposes of this article. First, ln α is linear in $1/F$ for several decades of α. Second, the same values of α are obtained for fields about 1·5 times larger in the recently studied 44-V junction as for the earlier 20-V junction. The first feature will be discussed in this subsection, and the second feature in subsection IIIC.

In this subsection we shall carry out the analysis in terms of

$$E_i^* = qF/\alpha \quad (2.14)$$

where q is the electronic charge and F/α is the voltage drop per ionization along the path of the ionizing carrier. Thus E_i^* is the average energy gained from the field by the carrier per ionization.

We shall consider two extreme limiting cases for α. One corresponds to very high values of F (very hot carriers) and the other to very low values of F.

In terms of the simple model for very hot carriers, the carrier energy is on the average greater than E_i. Then in order to produce an impact ionization the carrier will produce rE_R of phonon energy for each pair production which requires E_i. The high-field limiting value of E_i^* will thus be

$$E_i^* = E_i+rE_R \quad (2.15)$$

The low-field limiting form, derived below, is

$$E_i^* = rE_R\exp(E_i/qL_RF) \equiv rE_R\exp(b/F) \quad (2.16)$$

where

$$b \equiv E_i/qL_R \quad (2.17)$$

This low-field limit may be understood as follows. At low fields, the random energy of the carriers is small, in fact of the order of E_R. Consider a carrier scattered to an energy less than E_R. It will reach

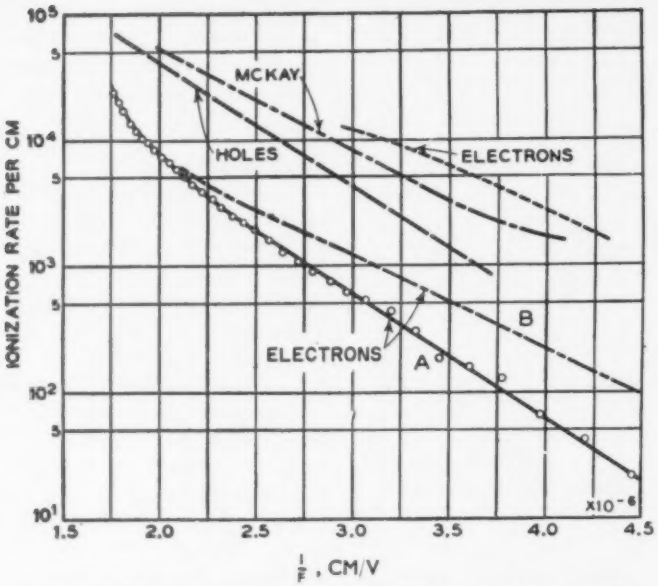


FIG. 6. Secondary multiplication coefficient in silicon according to CHYNOWETH[13]. Curve A corresponds to electrons in a junction of 44-V breakdown. The upper curve labeled "Electrons" applies to a previously measured 20-V breakdown junction. These are the two curves selected for analysis. (*B* represents an alternative interpretation of the data of *A*. The curve marked "holes" represents data reported by CHYNOWETH[13a] in the same publication as the highest curve. The remaining curve is from MCKAY[25].)

the energy E_i if it travels a distance E_i/qF without scattering. The probability of its doing this is only $\exp(-E_i/qL_RF)$. Thus a carrier must be scattered into an energy less than E_R a total number of times $r\exp(E_i/qL_RF)$ before it reaches ionization energy and then ionizes rather than produces a phonon. It must obtain energy E_R from the field for each of these events. Consequently E_i^* is given by (2.16) and α by

$$\alpha(F) = (qF/rE_R)\exp(-E_i/qL_RF) \qquad (2.18)$$

This treatment neglects ionization due to carriers which reach, for example, $E_i/2$, and then are scattered and subsequently gain the necessary energy. For small fields this neglect is justified because a carrier which has been scattered into a random direction with energy $E_i/2$ must on the average travel even farther to reach E_i than a carrier starting at rest. Thus, for small values of F, such two-stage processes are of negligible importance. The low-field limit holds, of course, only for fields so high that ordinary acoustical scattering energy losses can be neglected.

An intermediate situation may be estimated as follows: The law of conservation of energy requires that

$$qFv_d = (v_\theta/L_R)E_R^* \qquad (2.1)$$

where v_d is the drift velocity of carriers in the field and v_θ is a random hot carrier velocity, and E_R^* is the average energy loss per path length L_R. E_R^* varies from E_R for energies less than E_i up to $E_R+(E_i/r)$ for very hot electrons. The law of conservation of momentum similarly yields

$$qF = mv_dv_\theta/L_R \qquad (2.20)$$

The two equations may readily be solved for v_d and v_θ in the form

$$(1/2)mv_d^2 = E_R^*/2 \qquad (2.21)$$

$$E_\theta \equiv (1/2)mv_\theta^2 = (qL_RF)^2/2E_R^* = (E_i^2/2E_R^*)(F/b)^2 \qquad (2.22)$$

E_θ evidently corresponds to kT for the hot carriers.

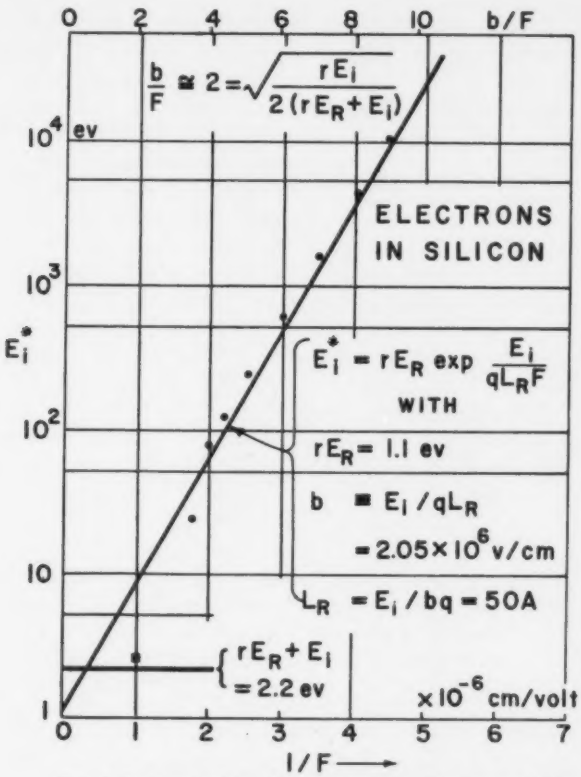


FIG. 7. Interpretation of Curve A of Fig. 6 in terms of the simple model using $E_i^* = qF/a$, the energy obtained from the electric field per pair produced. [The points ● are obtained from A of Fig. 6. The point ■ corresponds to the intermediate situation of equation (2.23).]

Similar expressions have been given by SHOCKLEY[1] and WOLFF[2]; WOLFF shows that the carrier distribution is approximately Maxwellian with kT replaced by E_θ but with the factor 2 replaced by $3\times$average $(1-\cos\theta)$ where θ is the angle of scattering. It is to this intermediate case that WOLFF fits the ionization data. He does not consider the low-field limit which appears to cover the important data discussed below.

From the E_θ expression it is seen that the very hot carrier case will be approached when E_θ is equal to E_i and

$$b/F \simeq (E_i/2E_R^*)^{1/2} \qquad (2.23)$$

At this condition about 7/8 of the carriers should be above E_i, and E_i^* should equal about (8/7) (E_i+rE_R).

The high-field, low-field and intermediate approximations are compared in Fig. 7 with CHYNOWETH's data A of Fig. 6. The low-field limit equation (2.16) with $\ln E_i^*$ varying as b/F holds very well over most of the range measured. For high fields the data deviate toward the intermediate case, shown at $b/F \simeq 2$ in accordance with equation (2.23) and $E_R^* = 0{\cdot}12$ eV. Fig. 7 leads to

$$b = 2{\cdot}05\times10^6 \text{ volts/cm} \qquad (2.24)$$

so that the remaining adjustable constant L_R is

$$L_R = E_i/qb = 5\times10^{-7} \text{ cm} \qquad (2.25)$$

It should be remarked that CHYNOWETH[13] has also interpreted b in the same way, but with $E_i = 2{\cdot}3$ eV. He does not, however, indicate how theory

might lead to this limiting expression, nor does he give any theory for the coefficient in front of the exponential.

Fig. 7 shows that the simple model with the three adjustable constants

$$E_t = 1{\cdot}1 \text{ eV}, rE_R = 1{\cdot}1 \text{ eV}, L_R = 50 \text{ Å} \quad (2.26)$$

now fits two additional experimental values, in addition to two from the enhanced quantum efficiency; namely the slope b and the intercept of E_i^* for $1/F = 0$. This intercept is in excellent agreement with the predicted value rE_R.

Actually the accurate agreement of the intercept value is probably fortuitous. As the paper of KELDYSH *et al.*[15] and the earlier work of FRANZ[16] shows, the effective energy gap is reduced in high electric fields. As a crude approximation, we assume that in the range of interest we may write

$$E_t(F) = (F_0 - F)E_t/F_0 \quad (2.27)$$

where F_0 is the field at which the gap would effectively vanish. Evidently F_0 is of the order of the field involved in tunnelling phenomena and is probably larger than b. Inserting this expression into the low-field form for E_i^* leads to

$$E_i^* = rE_R \exp(-b/F_0) \exp(b/F) \quad (2.28)$$

Thus the intercept should be expected to be smaller than rE_R by a factor of the order of $\exp(-1)$.

Table 1 shows the selected values for the constants of the present simple model together with corresponding values used in previous publications.

II E. *Average energy per hole–electron pair for high-energy particles*

When a high-energy particle such as a proton, β-ray, or even a fission fragment enters a silicon crystal, the number of hole–electron pairs created is independent of the nature of the particle and depends only on its energy, according to results recently reported by GRAINGER *et al.*[17] These results are in keeping with earlier results obtained by MCKAY and MCAFEE[12]. MCKAY and MCAFEE actually measured the energy required per pair (which apparently has not been done in the more recent results) and obtained $3{\cdot}6 \pm 0{\cdot}3$ eV per pair for silicon, and $2{\cdot}95 \pm 0{\cdot}15$ eV per pair for germanium.

Other values are given by VAVILOV[5] of $4{\cdot}2 \pm 0{\cdot}6$ eV for silicon and $3{\cdot}7 \pm 0{\cdot}4$ eV for germanium. ZAREBA[18] reports $2{\cdot}8 \pm 0{\cdot}3$ eV for germanium, but revises this later to $3{\cdot}8 \pm 0{\cdot}3$ eV[19].

The simple model interprets these results as follows. The high-energy particle dissipates its energy in the production of pairs of high-energy holes and high-energy electrons, which in turn can produce pairs, etc. The incident energy can be accounted for in three ways. Every pair-producing impact ionization consumes an amount of energy E_t from the kinetic energy of the system of carriers. In addition, in general, rE_R of energy is also converted to phonon energy before the ionization is produced. (It should be noted that most of the pairs are probably produced by carriers whose energy is only a few times E_t.) After a carrier has been produced and no longer ionizes, it will have kinetic energy of motion, denoted by E_f, and called the final energy which it dissipates to phonons.

It is evident that E_f will be of the order of magnitude of E_t since if a carrier is created just below energy E_t it will not be able to ionize and must dissipate E_t as heat. The same statement applies to a carrier which has reached E_t just after producing a hole–electron pair.

Table 1. *Secondary generation constants for* Si

	Optical or Raman		Ionization	
	Threshold	Mean free path	Threshold	Mean free path
	E_R	L_R	E_t	$L_t \equiv L_R/r$
Previous	0·07 eV	100–200 Å	2·3 eV (electrons)[14] 2·8 eV (holes)[14]	15 Å
Present	0·063 eV	50 Å	1·1 eV	880 Å

In terms of the final energy E_f the energy to produce a pair is

$$E_{\text{pair}} = 2E_f + E_i + rE_R \qquad (2.29)$$

If we assume that the carriers as finally produced are equally likely to be anywhere in the Brillouin zone with energy less than E_i, then for parabolic energy surfaces

$$E_f = 0{\cdot}6E_i \qquad (2.30)$$

and

$$E_{\text{pair}} = 2{\cdot}2E_i + rE_R \qquad (2.31)$$

This formula is in satisfactory agreement with the quoted measured values for E_{pair} as may be seen from Table 2.

Table 2. Energy in electron volts required per hole–electron pair produced by a high-energy particle or a high-energy photon

	Silicon	Germanium
E_i	1·1	0·66
rE_R	1·1	2·1
$2{\cdot}2\,E_i + rE_R$	3·5	3·6
E_{pair} measured:		
Ref. 18.	3·6 ± 0·3	2·95 ± 0·15
Ref. 5.	4·2 ± 0·6	3·7 ± 0·4
Ref. 19.		2·8 ± 0·3
Ref. 20.		3·8 ± 0·3

The simple model poses a statistical problem in regard to E_f which cannot be solved without further information on how the residual energy $E - E_i$ is distributed among the carriers subsequent to pair formation. A very crude treatment assuming that the energy is equally divided among the three carriers (incident carrier and pair) subsequent to pair formation leads to the result

$$E_f \simeq E_i(E_i + 3rE_R)/5(E_i + 0{\cdot}3rE_R) \qquad (2.32)$$

This does not change the silicon value of Table 2 and changes the germanium value to $E_f = E_i$ and $E_{\text{pair}} = 4{\cdot}1$ eV.

It is evident that within the accuracy of the data and the mathematical analysis, the simple model is in agreement with experiment for E_{pair}.

II F. *Interpretation of spectral distribution of radiation from a microplasma*

The radiation generated, presumably by recombination of holes and electrons in a microplasma (see Section V) has been measured by CHYNOWETH and MCKAY[20] and is shown in Fig. 8. This figure also shows radiation from a forward-biased junction.

It is to be noted that the microplasma radiation falls to about the same extent as the forward radiation for about seven times the energy increment. If the spread in energy is caused by kinetic energy of the carriers, this means a carrier temperature of about seven times room temperature or about 2100°K corresponding to $E_\theta = 0{\cdot}18$ eV.

This value of E_0 is somewhat smaller than that of the simple model which is

$$E_\theta = (qL_RF)^2/2E_R^* = 0{\cdot}32 \text{ eV} \qquad (2.33)$$

for $L_R = 50$ Å, $E_R^* = 0{\cdot}063$ eV and $F = 400{,}000$ V/cm; a reasonable field in a microplasma.

An alternative way to use the data is to solve the E_θ equation (2.22) for E_i. For this purpose we take WOLFF's more accurate form and take $3(1 - \cos\theta) = 3$ in place of 2. This gives

$$E_i = (b/F)(3E_\theta E_R^*)^{1/2} \qquad (2.34)$$

From Fig. 7 it is evident that breakdown fields, which probably are about 400,000–600,000 V/cm in microplasmas, correspond to b/F values of 3·3 to 5. The value of E_R^* can vary from 0·063 to 0·12 as discussed in subsection II D. This leads to a range of E_i values of

$$E_i = 1{\cdot}0 \pm 0{\cdot}4 \text{ eV} \qquad (2.35)$$

This is an estimate of the threshold for E_i based on secondary multiplication and plasma radiation. It is in good agreement with the value obtained from enhanced quantum yield.

Two other features of the radiation distribution should be noted. The curve continues with a single uniform behavior down to 1·1 eV. We shall return to this after discussing the high-energy portion of the curve. At high energies, the effective temperature appears to be lower since the curve is steeper. This is consistent with what might be expected due to the increase in E_R^* by about a factor of two with electron temperature.

The low-energy part of the curve is at first more puzzling. The tendency to turn upwards at the lowest energies shown may be due to transitions from one part to another of the complex energy-band structure. However, if recombination of

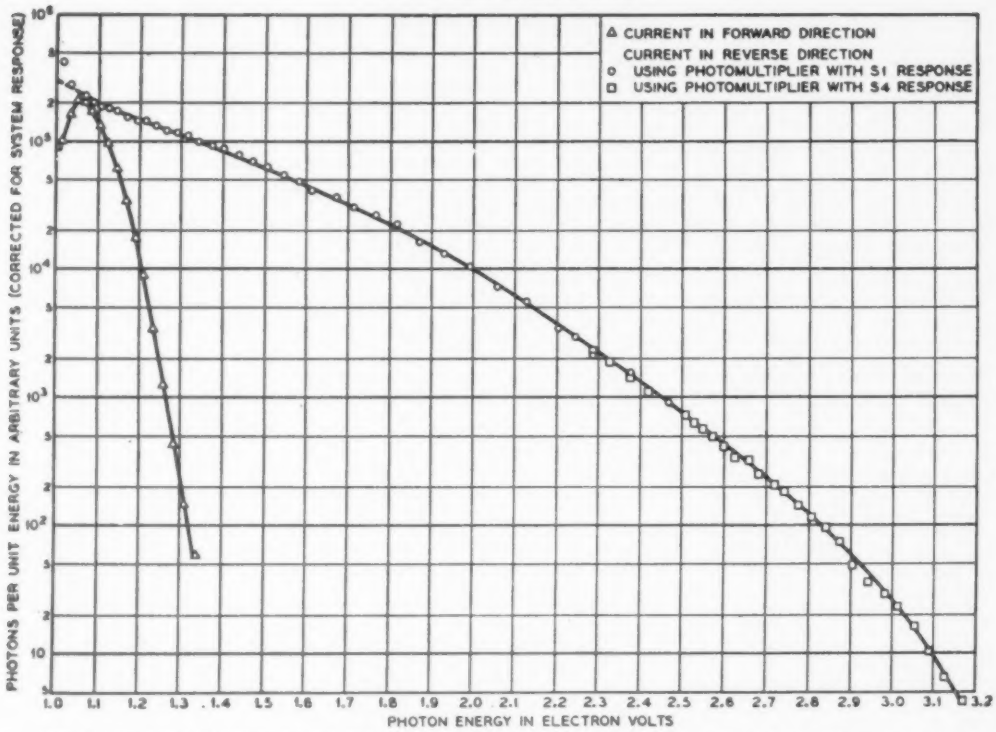

Δ CURRENT IN FORWARD DIRECTION
CURRENT IN REVERSE DIRECTION
o USING PHOTOMULTIPLIER WITH S1 RESPONSE
□ USING PHOTOMULTIPLIER WITH S4 RESPONSE
PHOTONS PER UNIT ENERGY IN ARBITRARY UNITS (CORRECTED FOR SYSTEM RESPONSE)
PHOTON ENERGY IN ELECTRON VOLTS
1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3.0 3.1 3.2

FIG. 8. Spectral distribution of photons per unit time per unit energy range from microplasma and from forward recombination (CHYNOWETH and MCKAY[20]).

hole–electron pairs produces most of the radiation, then the maximum should occur at an energy higher than the energy gap by a few times the random energy of equation (2.23). This maximum results from an expected spectrum of the form $E^n \exp(-E/E_\theta)$ where the E^n factor arises from the available space in the Brillouin zone times the probability of recombination. It is possible that this maximum is suppressed by a smooth merging of this contribution to the radiation with that from intraband transitions.

Another explanation, which appears more reasonable, is that in the high field of the microplasma the energy gap is effectively narrowed by the processes discussed in subsection II D. For the recombination mechanism discussed here, this means that a hole and electron which do not have enough energy to move together and recombine, tunnel together and thus recombine while their "dipole" energy in the field is negative, thus they radiate an energy less than the energy gap.

Returning to the high-energy end of the spectrum, we note that the decrease in intensity, or effective temperature, begins to occur for energies as low as 1·5 eV. If it is assumed that the intensity obeys a law of the form

$$I(h\nu) = E^n \exp(-E/E_\theta) \tag{2.36}$$

$$E = h\nu - E_g(F) \tag{2.37}$$

it is possible to choose a value of $E_g(F) = 0{\cdot}6$ eV and $n = 0{\cdot}5$ such that $I(h\nu)E^{-n}$ becomes straight up to $h\nu = 2$ eV. The high- and low-energy slopes then correspond to E_θ values of 0·27 and 0·17; the difference of a factor of about 1·5 is in keeping with low- and high-energy E_R^* values. The transition should occur when either the hole or the electrons has a substantial probability of requiring an energy greater than E_i. This should occur when $h\nu$ is somewhat less than $E_g(F)+2E_i = 2{\cdot}8$ eV and this is what is found from the experimental data. It

would appear more difficult to explain the data with a threshold of 2·3 eV.

It is evident that new and more extensive measurements of the plasma radiation, especially for uniform junctions like those discussed in Section IV, would be valuable.

Another interesting phenomenon which should give information on the electron temperatures is the electron emission from the surface of the silicon, observed by BURTON[21]. It has been impractical to try to interpret the data of BURTON or SENITZKY[22] for this study.

II G. *Interpretation of threshold data*

An extensive investigation of low-voltage-breakdown junctions has been carried out by CHYNOWETH and MCKAY[14] in an attempt to determine directly the threshold energy for pair production. By an empirical method of extrapolating data on photomultiplication, they conclude that in a junction of zero width no multiplication can occur unless the photocarrier falls through a drop of 2·3 V or more.

The narrowest junction they studied has a space-charge layer about 800 Å wide at the threshold. This is so wide compared to L_R, and the field is so low, that the low-field limit for secondary multiplication can be applied.

Fig. 9 shows an attempt to fit the data they use in their extrapolation. The fit is based on the values used for the simple model, and takes into account end effects for the narrow space-charge layer; however all values of α were arbitrarily multiplied by a factor of 4.

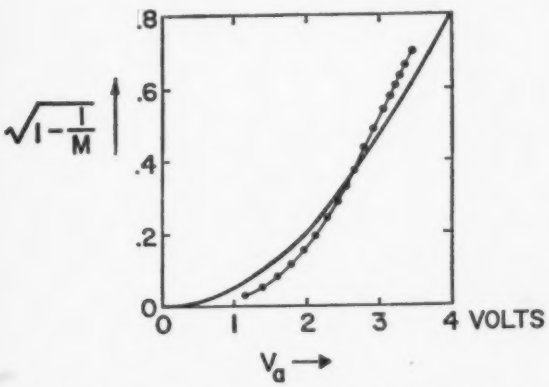


FIG. 9. A fit of the data used by CHYNOWETH and MCKAY[20] in their extrapolation to obtain the 2·3-eV threshold.

It is evident that the simple model with a threshold of 1·1 eV gives a curve of the observed shape with a straight-line portion which extrapolates to the value used by CHYNOWETH and MCKAY to obtain a 2·3 eV threshold. Thus it is evident that their data are not necessarily inconsistent with a threshold of 1·1 eV.

The factor of 4 needed to give the fit to the data may not be surprising in view of the low voltage and high degree of compensation in the junction (see Section III), and is less than the difference between CHYNOWETH's 44- and 20-V junction shown on Fig. 6.[13]

Thus, as stated in the introduction to this section, the simple model with three adjustable constants fits six experimental quantities: Threshold and slope for enhanced quantum yield in subsection II C; slope and intercept for low-field ionization coefficient in subsection II D; E_{pair} in subsection II E; and plasma radiation temperature in subsection II F.

III. SPATIAL FLUCTUATIONS (INHOMOGENEITIES) IN CHEMICAL CHARGE DENSITIES†

III A. *The statistics of the distributions of ions*

If donors and acceptors are produced by neutron capture in germanium, it is evident that if the expected number of donors in a volume U is

$$m = UN_d \tag{3.1}$$

then the probability of finding a donors in a volume U is given by the Poisson distribution

$$P(a) = m^a e^{-m}/a! \tag{3.2}$$

The average value of a is m and the mean-square deviation from the mean is

$$\text{av}\,(a-m)^2 = m \tag{3.3}$$

so that the root-mean-squared deviation of a from its mean value m, called the "deviation" hereafter and denoted by δm, may be written as

$$\delta m \equiv [\text{av}\,(a-m)^2]^{1/2} = m^{1/2} \tag{3.4}$$

The fractional deviation is evidently

$$\delta m/m = m^{-1/2} \tag{3.5}$$

† The treatment of this section follows and extends the previous work of the author.[23]

For large values of m the Gaussian approximation can be used:

$$P(a) = (2\pi m)^{-1/2} \exp -(a-m)^2/2m \quad (3.6)$$

It is evident that this factor would represent the entropy term $\exp(TS/kT) = \exp(S/k)$ in a statistical mechanical treatment of fluctuations.

When donors are not introduced by a purely random process but are allowed to diffuse at high temperature during the formation of a semiconductor by crystal growing from melt or solution, or by solid-state diffusion, electrostatic energy terms, in addition to entropy terms, will also tend to suppress their fluctuations. If a volume U contains a larger number a of donors than the expected number $m = UN_d$, then the electrostatic potential in this volume will be more positive, the increase in potential δV being that which will attract sufficiently more electrons and repel sufficiently more holes to neutralize the extra donors. (We are considering volumes large compared to a Debye length.) The change in electronic charge density produced by δV may be written as

$$-q(N_d-N_a)gq\delta V/kT = -q(a-m)/U \quad (3.7)$$

where it can readily be seen that

$$(N_d-N_a)g = 2n_i \text{ for } n_i > N_d-N_a \quad (3.8)$$

$$g = 1 \text{ for } n_i < N_d-N_a \quad (3.9)$$

$$g = 3kT/2(E_F-E_c) \text{ for deg. } n \quad (3.10)$$

where the last expression applies to parabolic bands for degenerate n-type.

To move $(a-m)$ donors into U against an average repulsive potential $\delta V/2$ requires an energy $\delta E(a)$.

$$\delta E(a) = q(a-m)\delta V/2 = [kT/2Ug(N_d-N_a)](a-m)^2 \quad (3.11)$$

This gives rise to a Boltzmann factor $\exp[-\delta E(a)/kT]$ tending to suppress the mean-squared fluctuations. Combining this with the entropy factor represented by the Poisson distribution leads to

$$P(a) = A \exp -\frac{(a-m)^2}{2m}\left[1+\frac{N_d}{g(N_d-N_a)}\right] \quad (3.12)$$

where the Gaussian approximation is used and A is a normalization factor.

For the cases of chief interest, corresponding to high-temperature intrinsic conditions

$$g(N_d-N_a) = 2n_i \gg N_d \quad (3.13)$$

so that the Boltzmann term has little effect in suppressing the fluctuations.

Thus in the following we shall assume that the chemical charges of donors and acceptors are distributed statistically and independently in accordance with a Poisson distribution.

A consequence of this conclusion is that donors and acceptors fluctuate independently so that the root-mean-square fluctuation δQ of chemical charge in a volume U is given by

$$(\delta Q)^2 = q^2U(N_d+N_a) = Cq^2U(N_d-N_a) \quad (3.14)$$

where

$$C \equiv (N_d+N_a)/(N_d-N_a) \quad (3.15)$$

is the "compensation factor" and is the ratio of total chemical charge density to unbalanced charge density. The effect of compensation on the Gaussian approximation to the Poisson distribution is to replace $m = |U(N_d-N_a)|$ by Cm where it affects the spread of the distribution in expressions like equation (3.4) for δm or in equation (3.6) where Cm should replace m in the normalization coefficient and in the denominator of the exponential term.

III B. *Application to the distribution of breakdown voltage in an abrupt p^+–n junction*

The condition for avalanche breakdown, which we shall discuss more fully in the next two sections, can be approximated by the condition that the peak field reach a breakdown value F_B. In an abrupt or step junction (p^{++})–n this leads to

$$V_B = F_BW_B/2 \quad (3.16)$$

$$KF_B = q(N_d-N_a)W_B \quad (3.17)$$

$$V_B = KF_B/2q(N_d-N_a) \quad (3.18)$$

where W_B is the space-charge width, and V_B the voltage at breakdown and

$$K = 1{\cdot}04\times10^{-12} \text{ farad/cm} \quad (3.19)$$

is the dielectric constant for silicon.

The peak field occurs at the junction and its fluctuations at any point will be largely determined by fluctuations in ion density in the immediate

neighborhood. We shall use a "cube assumption" and divide the space-charge layer into cubes and suppose that in each cube the breakdown is determined by the average ion density in a cube of edge W_B containing on the average

$$m = (N_d - N_a)W_B^3 = 4V_B^2/(qF_B/K) \simeq 50\,V_B^2 \tag{3.20}$$

The quantity

$$qF_B/K \simeq 0{\cdot}08 \text{ volt}^2 \simeq (0{\cdot}3 \text{ volt})^2 \tag{3.21}$$

is a characteristic mean-squared fluctuation voltage in silicon. It determines the scale of spread of voltage over the junction.

Since the voltage V_B depends, according to equation (3.18), on the reciprocal of chemical charge density, it follows that $\delta V_B/V_B$, the fractional root-mean-squared fluctuation of V_B from cube to cube, is the same as for chemical charge density, and hence the same as the fractional root-mean-squared fluctuation of chemical charge in W_B^3:

$$\begin{aligned}\delta V_B/V_B = \delta Q/Q &= (Cm)^{1/2}/m \\ &= [(N_d+N_a)W_B]^{1/2}/(N_d-N_a)W_B^3 \\ &= [C^{1/2}(qF_B/K)^{1/2}]/2V_B \end{aligned} \tag{3.22}$$

where C is the compensation factor of equation (3.15). From this it follows that the deviation in voltage is practically independent of V_B and is given by

$$\delta V_B = C^{1/2}(qF_B/K)^{1/2}/2 \simeq C^{1/2}0{\cdot}14 \text{ V} \tag{3.23}$$

The foregoing results are summarized on Fig. 10.

A typical junction of, say, 20-V breakdown has $W_B = 0{\cdot}4\,\mu$ and an area of 10^{-3} cm^2 $= 10^5\,\mu^2$. It contains about $N = 10^6$ cubes each containing, according to equation (3.20) about $m = 20{,}000$ unbalanced ions. With so many cubes, it is evident there there will be some which break down much lower than the average.

We now ask what will be the breakdown voltage for the cube of lowest breakdown voltage? This "extreme" cube will contain the largest number of ions of any cube in the junction.

The probability that a cube will contain b or more ions is the sum $S(b, m)$ where

$$\begin{aligned} S(b, m) &= \sum P(a, m) \text{ for } a \geqslant b \\ &\simeq (Cm/2\pi)^{1/2}(b-m)^{-1} \exp -(b-m)^2/2Cm \end{aligned} \tag{3.24}$$

The last expression follows from the Gaussian approximation with the effect of compensation included according to equation (3.15). The

$$V_B = \frac{F_B W}{2} = \frac{F_B^2 K}{2qN_d} = \frac{F_B^2 K \mu_n \rho_n}{2} \cong 100\rho_n$$

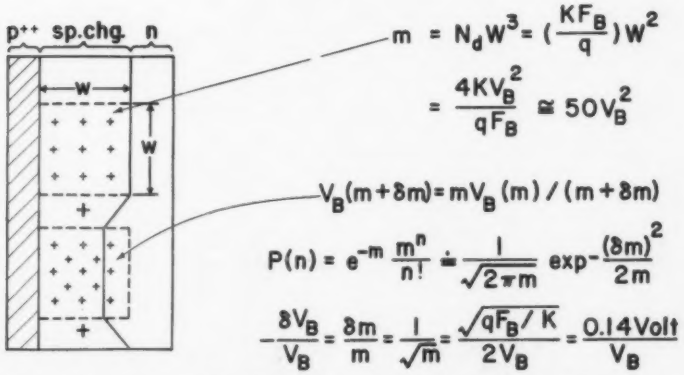

p++ sp.chg. n
w
w
$m = N_d W^3 = \left(\frac{KF_B}{q}\right) W^2$
$= \frac{4KV_B^2}{qF_B} \cong 50V_B^2$
$V_B(m+\delta m) = mV_B(m)/(m+\delta m)$
$P(n) = e^{-m}\frac{m^n}{n!} \doteq \frac{1}{\sqrt{2\pi m}} \exp -\frac{(\delta m)^2}{2m}$
$-\frac{\delta V_B}{V_B} = \frac{\delta m}{m} = \frac{1}{\sqrt{m}} = \frac{\sqrt{qF_B/K}}{2V_B} = \frac{0.14 \text{ Volt}}{V_B}$

FIG. 10. Method of estimating the fluctuation of breakdown voltage from one region to another of an abrupt $(p^+)n$ junction using the "independent cubes" assumption. (To be consistent with the text, δm should be replaced by $a-m$ except in the last line.)

probability that a cube contains $b-1$ or less ions is then

$$S'(b, m) = 1-S(b, m) \tag{3.25}$$

The probability that all the N cubes simultaneously contain $b-1$ or less ions is then

$$[S'(b, m)]^N = [1-S(b, m)]^N \simeq$$
$$\simeq \exp -[NS(b, m)] \equiv f \tag{3.26}$$

(The approximate equality is very precise for large N.) The quantity f is evidently the fraction f (a function of b, m and N) of the junctions which have no cube with more than $b-1$ ions. Therefore, f is the fraction of the junctions which break down above $V_B(b)$, that is, break down at $V_B(b-1)$ or higher. The fraction of the junctions which have extreme cubes with precisely b ions is evidently $f(b+1, m, N)-f(b, m, N)$, which we can, but shall not, express in terms of the derivative in respect to b of $S(b, m)$.

In order to obtain an idea of the distribution from junction to junction of $V_B(b)$ corresponding to the extreme low-breakdown cube, it suffices to consider three values for f, namely $f = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$. Further treatments of the statistics of extremes will be found, for example, in GUMBEL[24].

Half the junctions have an extreme b value no larger than that given by solving equation (3.24) for $S(b, m)$ with $f = \frac{1}{2}$. This value of b then represents the junction of median breakdown voltage. Letting f be general for the moment, we obtain

$$S(b, m) = -(\ln f)/N \tag{3.27}$$

and using the Gaussian approximation we obtain

$$b-m = \{2Cm[\ln N - \ln(-\ln f)$$
$$-\tfrac{1}{2}\ln 2\pi(b-m)^2/2Cm]\}^{1/2} \tag{3.28}$$

For $N \simeq 10^6$ and $m \simeq 10^4$, the value of $(b-m)^2/2\,Cm$ is approximately 15, so that the $(b-m)^2/2\,Cm$ term on the right side is only about 15 per cent of $\ln N$. Therefore, near the median [$f = \frac{1}{2}$ and $\ln(-\ln f] = -0{\cdot}037$s of the extreme distribution

$$b-m = (2Cm \ln N)^{1/2} \tag{3.29}$$

with an accuracy of better than 10 per cent.

The spread of the extremes can be estimated by considering a range of f from $\frac{1}{4}$ to $\frac{3}{4}$ which contains 50 per cent of the junctions. Actually the distribution is unsymmetrical, GUMBEL[24]. The change in $\ln(-\ln f)$ is about 1·6 and the change in $(b-m)$ is 0·8 $(2\,Cm/\ln N)^{1/2}$ so that the mean deviation of $(b-m)$ from its median value is approximately

$$\delta m_{\text{extreme}} \simeq (Cm/\ln N)^{1/2} \tag{3.30}$$

where $\delta m_{\text{extreme}}$ is somewhat arbitrarily defined as

$$\delta m_{\text{extreme}} = \text{change in } (b-m) \text{ from } f = \tfrac{1}{4} \text{ to } f = \tfrac{3}{4} \tag{3·31}$$

Thus $\delta m_{\text{extreme}}$ is smaller than δm by the factor $(\ln N)^{-1/2}$. Hence the deviation in voltage for the extreme low-voltage cube from junction to junction is

$$\delta V_{B\,\text{extreme}} = \delta V_B/(\ln N)^{1/2}$$
$$= (C/\ln N)^{1/2}(qF_B/K)^{1/2}/2$$
$$\simeq C^{1/2}0{\cdot}04 \text{ volt} \tag{3.32}$$

the last form corresponding to $N = 10^6$.

From equation (3.29) it follows that a typical junction will have its extreme low-breakdown cube lower in voltage by $(2 \ln N)^{1/2}$ times the deviation δV_B from cube to cube. For $N = 10^6$ this factor is about 5 so that

$$V_{B\,\text{extreme}} \simeq V_B(m)-5\delta V_B$$
$$\simeq V_B(m)-0{\cdot}7C^{1/2} \text{ volt} \tag{3.33}$$

There will be a rapidly increasing number of cubes with breakdown voltage slightly above $V_{B\,\text{extreme}}$. This distribution may be estimated from the Gaussian approximation which, including compensation, varies as

$$P(a, m) \propto \exp(b-m)(b-a)/Cm$$
$$= \exp V_1(2m \ln N)^{1/2}/C^{1/2}V_B \tag{3.34}$$

where V_1 is the increment of breakdown voltage of a over the low extreme $V_{B\,\text{extreme}}$:

$$V_1 = V_B(a)-V_{B\,\text{extreme}} = (b-a)V_B/m \tag{3.35}$$

Thus the slope of ln (number of cubes that can break down) versus V_1 (voltage above first break-

down) should be

$$(2m \ln N)^{1/2}/C^{1/2}V_B \simeq \simeq (8 \ln N/C)^{1/2}(K/qF_B)^{1/2} \simeq \simeq [(8 \ln N)/C]^{1/2}3 \text{ volt}^{-1} \simeq \simeq 30C^{-1/2} \text{ volt}^{-1} \quad (3.36)$$

the last expression applying to $N = 10^6$. We shall return to this expression in discussing "uniform" breakdown junctions in Section V.

III C. *The effective local fluctuations on the secondary multiplication coefficient*

A carrier, moving under the influence of an average electric field F in the space-charge layer of a p–n junction may be aided by statistical fluctuations in voltage in acquiring the necessary energy E_i for ionization discussed in Section II. These fluctuations will be larger the higher the degree of compensation. They also have a bigger fractional effect upon the field in lower-breakdown-voltage junctions.

As a consequence of this, at the same average field the ionization rate will be higher in a low-voltage junction. This is in the direction observed by CHYNOWETH[13] and discussed in connection with Fig. 6. In this section we shall estimate the magnitude of the fluctuation effects and compare the results with experiment.

We shall deal with the low-field limit for the simple model of ionization, subsection II C. We consider a carrier scattered at a point P_1. It is accelerated in the direction of the field F and moves a distance x. At x it has gained an energy qFx from the field. In addition, about 1/4 of the time it has gained an additional energy of $q\delta V(x)$ where

$$\delta V(x) = \text{root-mean-square fluctuation in voltage between two points separated by } x \quad (3.37)$$

If this gain in energy is greater than the energy E_i required to ionize, which we write as

$$E_i \equiv qV_i \quad (3.38)$$

then the chance of ionization is L_R/L_i.

We shall first calculate the mean-square fluctuation $[\delta V(x)]^2$. For this purpose we consider an element of volume dU which contains $(N_d+N_a)dU$ ions. The mean-square fluctuation of charge as discussed in equation (3.14) is given by

$$d(\delta Q)^2 = q^2(N_d+N_a)\,dU \quad (3.39)$$

This produces a contribution to the fluctuation of voltage, calculated from the coulomb potential, $q/4\pi Kr$, given by the expression

$$d[\delta V(x)]^2 = [q^2/(4\pi K)^2] \times \times [(1/r_1)-(1/r_2)]^2(N_d+N_a)\,dU \quad (3.40)$$

where r_1 and r_2 are distances of dU from the points separated by x. Since each element of volume fluctuates independently, this expression may be integrated. (It is convenient to integrate terms of the form $r_1{}^2-(r_1r_2)^{-1}$ in spherical polar coordinates about $r_1 = 0$.) The result for a uniform distribution of (N_d+N_a) is

$$[\delta V(x)]^2 = (q^2/4\pi K^2)(N_d+N_a)x \quad (3.41)$$

We now re-express this for a linearly graded junction which at breakdown has chemical charge density at the edge of the space-charge layer denoted by N_B. We assume N_B is small compared to the total ion density (N_d+N_a), which is taken as substantially constant. We introduce a compensation factor C

$$C = (N_d+N_a)/N_B \quad (3.42)$$

Other necessary relationships are

$$V_B = 2F_BW_B/3 \quad (3.43)$$

$$KF_B = qN_BW_B/4 \quad (3.44)$$

where F_B is the peak field at the center of the junction when the breakdown voltage V_B is applied and the space-charge width is W_B.

In terms of these quantities

$$[\delta V(x)]^2 = (C/\pi)(qF_B/K)(x/W_B) = (2/3\pi)(qF_B/K)(F_Bx/V_B) \equiv \beta^2x \quad (3.45)$$

This expression will be valid only for $x/W_B \ll 1$ since edge effects were disregarded in integrating over dU to obtain $[\delta V(x)]^2$.

The statistical problem of calculating the effect of the fluctuations on the probability of ionization is straightforward, but too complex to deal with exactly here. Consequently, a simplified approximation is employed. In the absence of fluctuations,

ionization can first occur after a distance

$$x = E_i/qF \equiv V_i/F \tag{3.46}$$

At this distance the fluctuation in voltage is $\beta x^{1/2}$. Thus the carrier need travel only a fraction

$$(V_i - \beta x^{1/2})/V_i = 1 - \beta x^{1/2}/V_i \tag{3.47}$$

of the uniform field distance in order to have approximately a 25 per cent chance of having the ionization energy $E_i = qV_i$. This is roughly equivalent to having an effective field $F^\dagger$ which is larger than F by about

$$F^\dagger/F = [1 + (F/b)\ln 4 - (F_f/F)^{1/2}]^{-1} \tag{3.48}$$

(The term in ln 4 is unimportant for small fields.)

For example, a ratio of 3/2 for $F^\dagger/F$ is obtained for $F_f/F = 1/9$. If this ratio occurs when $F = F_B/3$ then $F_f/F_B = 1/27$ and F_f is about 20,000 V/cm. The corresponding value of C is

$$C = 150 V_B F_f/F_B = 5{\cdot}5 V_B \tag{3.51}$$

which gives a compensation ratio of 100 for a 20-V junction.

Fig. 11 shows the data from Fig. 6 corrected by correcting the $1/F$ values for the 20-V junction by the formula

$$1/F^\dagger = (1/F)[1 - (F_f/F)^{1/2}] \tag{3.52}$$

with $F_f = 17{,}000$ V/cm.

The possible importance of compensation and

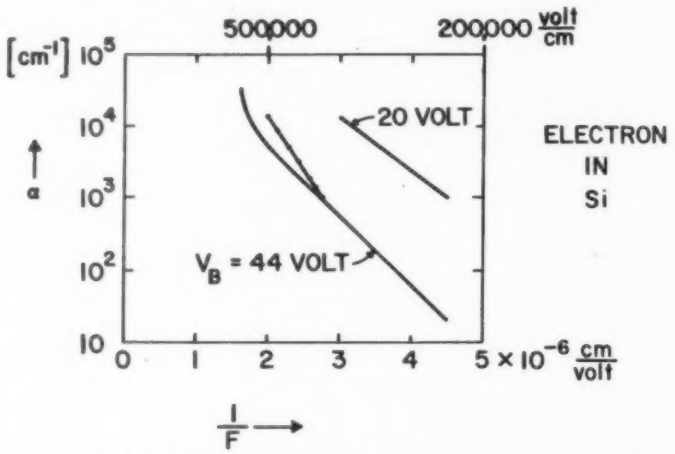


FIG. 11. Correction of CHYNOWETH's 20-V breakdown data using a fluctuation field of $F_f = 17{,}000$ V/cm.

The fluctuation field F_f is defined by writing

$$\beta^2 x/V_i^2 = F_f/F = \beta^2/V_i F = (2/3\pi)C[(qF_B/K)/(V_i V_B)](F_B/F) \tag{3.49}$$

so that

$$F_f = \beta^2/V_i = (2/3\pi)C[(qF_B/K)/(V_i V_B)]F_B \simeq \simeq (C/150V_B)F_B \tag{3.50}$$

the last form corresponding to taking $V_i \simeq 1{\cdot}1$ eV. The effect of the term in F_f/F in $F^\dagger/F$ can lead to discrepancies as large as those discussed in connection with CHYNOWETH's 44- and 20-V junctions.

statistical fluctuations does not seem to have been appreciated. Consequently, information which permits an estimate of C is not generally quoted in the literature pertaining to secondary ionization. We can, therefore, make only general observations in support of the model proposed here to account for variations in effective field values.

If, in a uniform field F, the secondary multiplication factor is $\alpha(F)$, then in a fluctuating field with average value F, α will have approximately the dependence $\alpha^\dagger(F)$ where

$$\alpha^\dagger(F) = \alpha(F^\dagger) = \alpha\{F/[1 - (F_f/F)^{1/2}]\} \tag{3.53}$$

The effect of this dependence will be to make measured values of α higher than the true α and less dependent on field, since $F^\dagger$ does not decrease as rapidly with field as does F itself. This is precisely the effect observed by McKay and McAfee[12] and described by Chynoweth[13] who points out that the curve for an individual junction shows less dependence upon F than a composite curve based on a set of different junctions.

It should also be noted that a "fine-textured" effect is the most natural way to explain a systematic shift of effective field. Variations on a large scale, such as the cube-to-cube variations discussed in subsection III B, would give rise to patchy behavior like that discussed in the next section. The distance over which the fluctuations are important is of the order of $E_i/qF = V_i/F$ or 500 Å at a field of 200,000 V/cm. There will be, in general, about V_B/V_i such regions across the junction, and thus a large amount of averaging over the various fluctuations can occur even in the passage of a single carrier across the junction. This can result in an effective increase in ionization rates just as if the effective field were raised from F to $F^\dagger$.

IV. THE $1/M(V)$ VERSUS V TEST OF PATCHINESS

The fact that gross fluctuations from region to region in p–n junctions have not played an important role in determining pre-breakdown multiplication characteristics can be shown in terms of a "patchiness" test. In order to understand this test the dependence upon voltage of the multiplication factor $M(V)$ for photocurrent or other injected currents must be understood.

As various authors have shown[25,26,13]

$$1-1/M(V) = \int_0^W \alpha[F(x)]\,dx = m\alpha(F)W \equiv S(V) \tag{4.1}$$

where $\alpha(F)$ is an appropriate average of the secondary multiplication coefficients of holes and electrons evaluated at the peak field F which occurs at the center of the space-charge layer of the graded junction considered in subsection III C. Chynoweth finds that the integral is accurately represented by taking $m = 0{\cdot}32$. The integral is in effect the ratio of the charges carried by successive generations of transits across the junction, $M(V)$ being the sum $1+S+S^2+\ldots = (1-S)^{-1}$. (Wul and Shotov[27] have given a more exact treatment than (4.1) for the case in which the coefficients for holes and electrons have a constant ratio.)

Breakdown occurs when $M(V)$ approaches infinity or S approaches unity. Since $S(V)$ has no singularity at $V = V_B$ where

$$S(V_B) = 1 \tag{4.2}$$

evidently near V_B

$$1/M(V) = [dS(V)/dV]_{V_B}(V_B-V) - 1/2[d^2S(V)/dV^2]_{V_B}(V_B-V)^2+\ldots \tag{4.3}$$

Thus $1/M(V)$ should approach zero linearly as V approaches V_B. An index n may be introduced to describe the slope by writing

$$1/M(V) = 1-S = [1-(V/V_B)]n = (1-y)n \tag{4.4}$$

$$y \equiv V/V_B \tag{4.5}$$

so that when $V = V_B$ and $y = 1$

$$V_B[dS(V)/dV] = dS(y)/dy = n \tag{4.6}$$

Miller[26] has found the expression

$$M(V) = 1/[1-(V/V_B)^n] \tag{4.7}$$

useful in fitting multiplication data. This corresponds to taking

$$S(V) = (V/V_B)^n \tag{4.8}$$

and

$$V_B[dS(V)/dV]_{V_B} = n \tag{4.9}$$

A theoretical value of n can be given in terms of the low-field limit for $\alpha = qF/E_i^*$ given by equation (2.18) with an effective average $\bar{b}$ for holes and electrons and F_B and F_f taken as defined in subsection III C. The dependence of $S(V)$ upon y may then be seen to be of the form

$$S(y) = mW\alpha^\dagger(F) = qmWF^\dagger/\overline{rE_R}\exp\bar{b}/F^\dagger \tag{4.10}$$

$$1/F_{\max}^* = [1-(F_f/F)^{1/2}]/F \tag{4.11}$$

where rE_R^* is an appropriate average for holes and electrons. Noting that WF is proportional to y and

$$F/F_B = y^{2/3} \tag{4.12}$$

straightforward manipulation, letting

$$(F_f/F_B)^{1/2} \equiv u \tag{4.13}$$

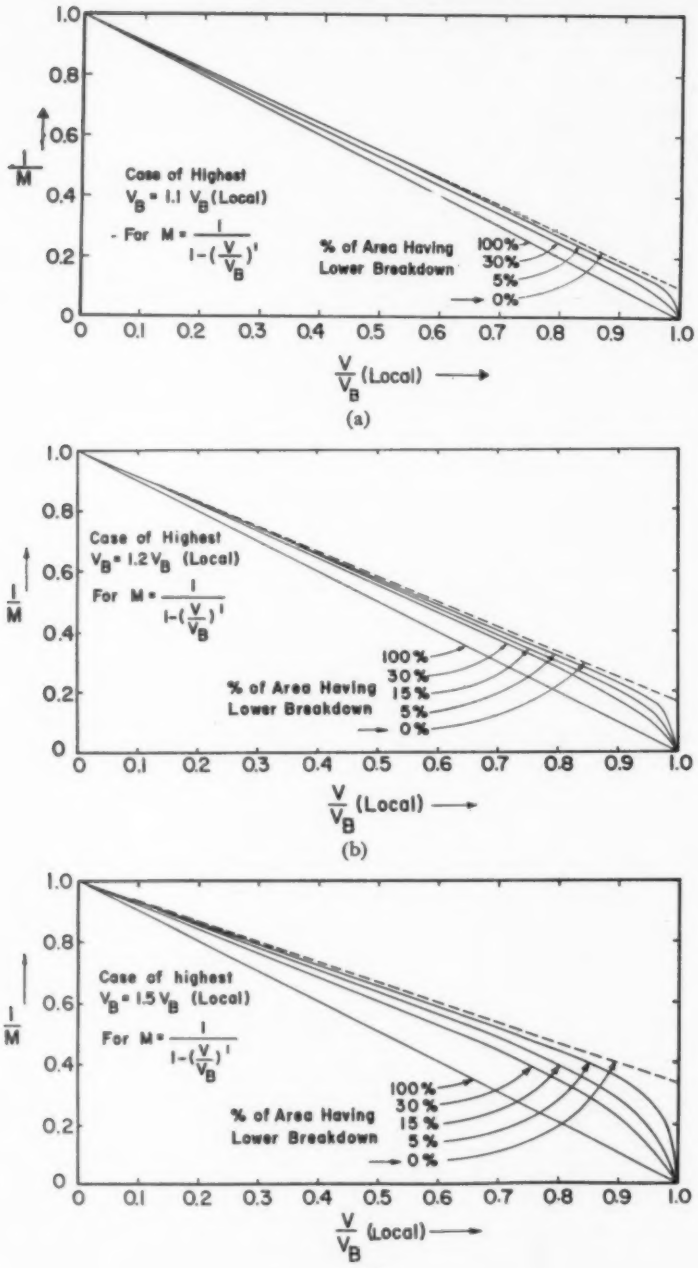


FIG. 12. Calculated curves for $I/M(V)$ versus V/V_B for "patchy" junctions having a certain fraction of their area at a lower breakdown voltage than the remainder; the higher breakdown voltage being larger by the following factors: (a) 1·1; (b) 1·2; (c) 1·5.

leads to

$$n = 1-[u/3(1-u)]+(\bar{b}/F_B)[(2/3)-u]$$
$$= 1-[u/3(1-u)]$$
$$+[(2/3)-u][1-u]^{-1}\ln[qV_B/rE_R^*(1-u)] \quad (4.14)$$

where $3m/2$ has been taken to be $1/2$ in relating V_B to $\bar{b}/F^\dagger$ at $S = 1$. This formula leads to values of n which are generally in agreement with values reported by MILLER[26]. However, extensions of both the data and the theory would be required before an adequate test of the model could be made.

If a junction consists of several regions of different breakdown voltage, then the observed breakdown should be determined by the regions of lowest breakdown voltage unless these are very small. If the low-breakdown regions are very small, series-resistance effects may limit their currents. (Such regions have been studied by BATDORF *et al.*[28] and by GOETZBERGER and SHOCKLEY[29]).

If a junction consists of a relatively small proportion of relatively low breakdown patches in an area of higher breakdown, then the effective n values defined, in keeping with equation (4.9), as

$$n = -V_B[dM(V)^{-1}/dV]_{V_B} \quad (4.15)$$

will be drastically increased because the low-breakdown areas become important as their breakdown voltage is approached.

Fig. 12(a–c) shows calculated $1/M(V)$ versus $V/V_B = x$ for various patchy junctions computed for $n = 1$ for the patches. As x approaches unity, the slope approaches the reciprocal of the fraction of the area at lower breakdown. Thus if, for example, only 10 per cent of the area is of lower breakdown, the apparent n values will be raised by a factor of 10. In general

$$n \text{ (patchy case)} = \frac{n \text{ (uniform area)}}{\text{(fraction of area that has lowest breakdown)}} \quad (4.16)$$

Pronounced curvature near $x = 1$ is also noted.

Fig. 13(a–e) shows experimental data obtained by various authors replotted in the form $1/M(V)$ versus V or V/V_B. It is seen that these curves are in general quite straight near V_B and correspond to reasonable n values. Thus we conclude that patch effects are playing only a small role in affecting the multiplication characteristics when $M = 10$ or less.

This evidence that patchiness is not important supports the view of subsection III C that the phenomenon which gives rise to different effective fields in different junctions is of a very fine-grained character and spread uniformly over the junction. This interpretation disagrees with BATDORF *et al.*[28] who propose that effective field differences are caused by localized effects such as microplasmas. These observations further support the view that fluctuations enhanced by compensation may cause the difference in ionization coefficients for the 44- and 20-V junctions discussed in subsection III C.

V. MICROPLASMAS AND RELATED MATTERS

V A. *Microplasma phenomena in general*

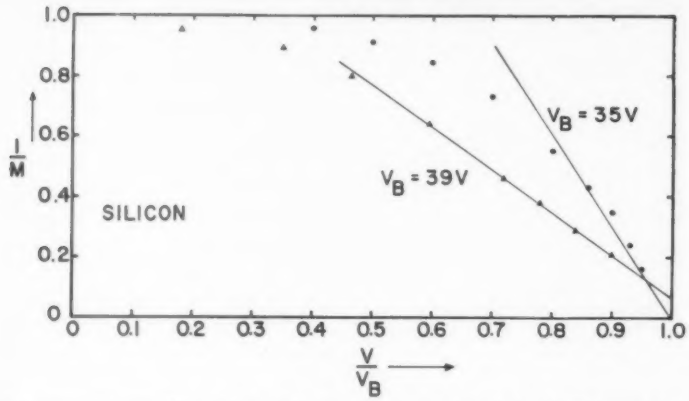
The breakdown of p–n junctions in silicon is usually characterized by microplasma effects. The earliest observations were made by MCKAY[26] and MCKAY and MCAFEE[12], and have been continued by MCKAY and his colleagues, particularly in publications by CHYNOWETH and MCKAY[20,31]

In brief, it has been found that the current is a sort of intermittent step function of the voltage, being either zero or the order of 30–100 μA when the current source turns on. We shall refer to this behavior as an "on-off" current characteristic. "On-off" characteristics produce microplasma noise which has been studied by the authors quoted and also by CHAMPLIN[32].

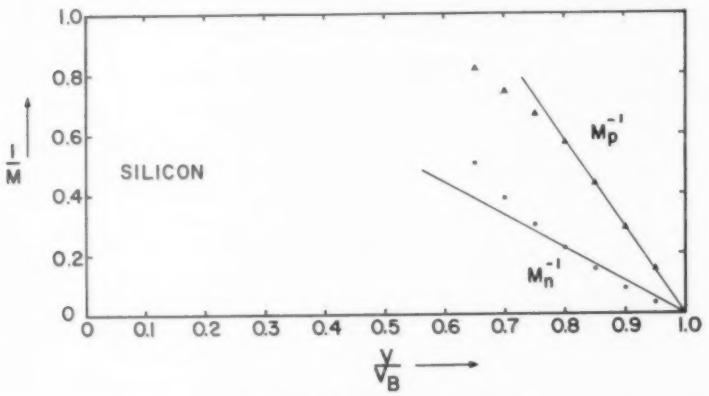
The "on-off" current sources have a duty cycle that varies with the applied voltage and gives rise to the variation in average current.

A slight negative resistance may be associated with the "on–off" microplasma, and the voltage across the p–n junction may drop by a fraction of a volt when the microplasma turns on. These effects have been studied by SENITZKY and MOLL[33] and others.

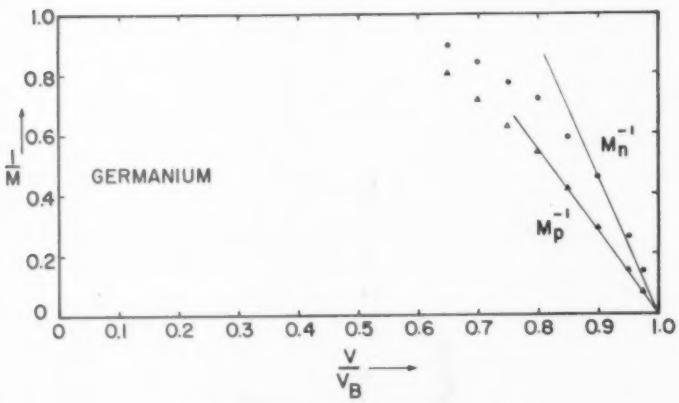
ROSE[34] proposed a model for a microplasma based on analogy with gas-discharge phenomena to explain microplasma effects. Fig. 14 shows the field distribution he considers. It is seen that in the "on" condition the field distribution bears almost no resemblance to the "off" condition. Con-



(a) Silicon (MILLER[26]). (The 35-V junction is n–p^+ and the 39-V junction is p–n^+.)

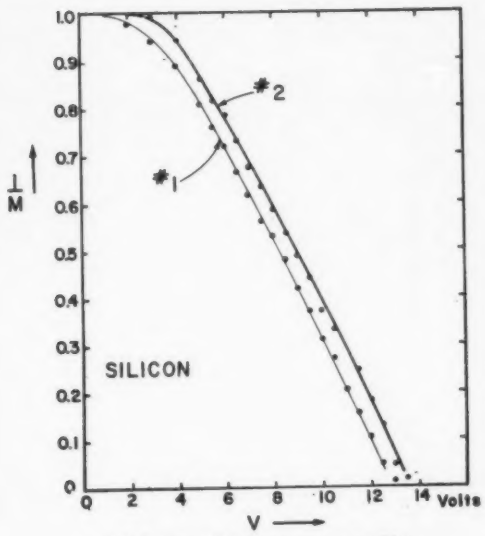


(b) Silicon (WUL and SHOTOV[27]).

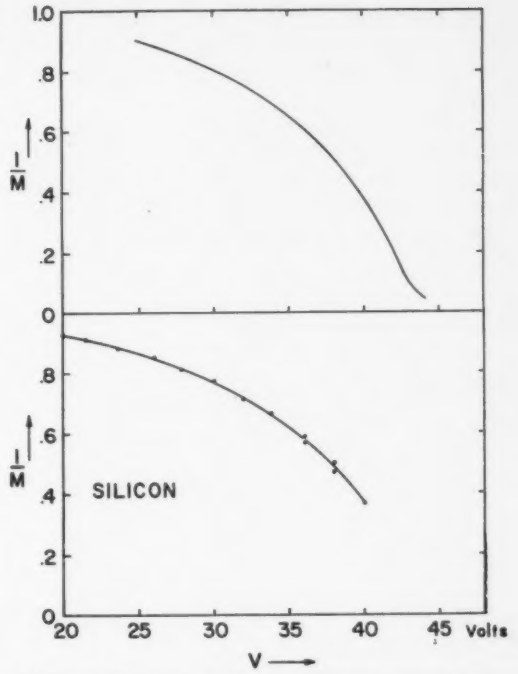


(c) Germanium (WUL and SHOTOV[27]).

FIG. 13. Plots of experimental data in the form $I/M(V)$ versus V/V_B.



(d) Silicon (McDONALD *et al.*[30]).



(e) Silicon (BATDORF *et al.*[28] The upper curve is from their Fig. 6 and the lower from Fig. 8).

FIG. 13 (*contd.*).

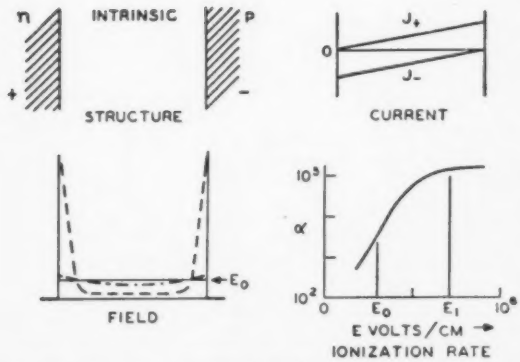


FIG. 14. Diagrams illustrating the mechanism of instability leading to a microplasma for an n-intrinsic-p structure, after ROSE[34].

sequently, the voltage in the "on" condition may be only a fraction of that in the "off" condition.

ROSE's model employs a p-intrinsic-n structure. As he points out, space-charge conditions will be very different in a linearly graded junction. In the next subsection we shall discuss this problem and come to the conclusion that noisy microplasmas with "on–off" current characteristics should not form in linearly graded p–n junctions.

Evidence that effects like those ROSE has considered may occur in suitable geometries is furnished by the work of GIBSON and MORGAN[35]. Fig. 15 shows the switching characteristics of one of their diodes which apparently contains plasmas similar to those considered by ROSE.

It is the present writer's opinion, and also a view emphasized by BATDORF *et al.*,[28] that microplasmas involve structural defects and would not occur in junctions containing only statistical fluctuations like those considered in Section III.

It would appear that microplasmas require structural imperfections having the following two features:

(1) The microplasma-producing structural imperfection produces a local electric field disturb-

ance over a region in which the voltage drop is only about 1 V (i.e. the order of E_i/q). Thus the region is only a few hundred Angstroms in length along the field.

(2) The structural imperfection contains traps which can immobilize high densities of charge. These charges set up electric fields comparable to F_B and provide a "lock-on" mechanism which gives rise to the "on–off" characteristic of the microplasma.

breakdown in structurally perfect p–n junctions is also presented.

In subsection V C, precipitated quartz particles are proposed as causes of microplasmas, and the role of dislocations is briefly discussed.*

V B. *Breakdown theory for structurally perfect p–n junctions*

In this subsection we shall give a treatment of the behavior to be expected when the breakdown

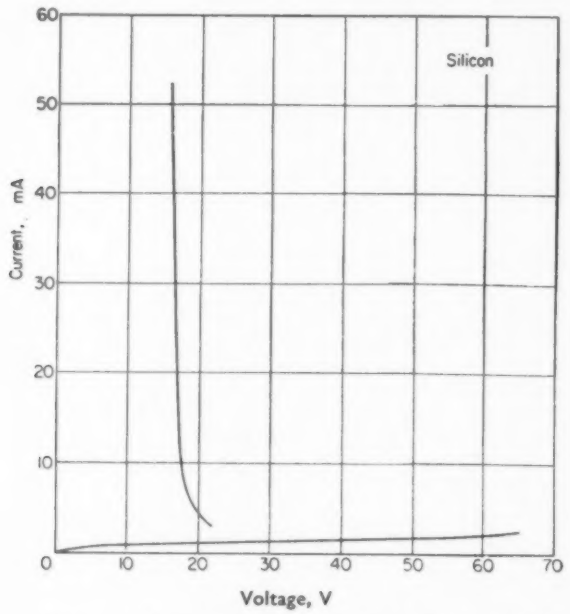


FIG. 15. Current–voltage characteristic of a typical silicon avalanche diode of the form studied by GIBSON and MORGAN[35].

The necessity for the first requirement is the experimental observation that the voltage across the junction changes by only a small amount when the microplasma switches on. If the microplasma extended across much of the space-charge layer, the change in voltage should, in keeping with Figs. 14 and 15, be a substantial fraction of the breakdown voltage V_B rather than a fraction of a volt.

The second requirement follows from the fact that space-charge of mobile carriers will in general tend to stabilize the microplasma by contributing a high series resistance. This point is further discussed in the next subsection. The theory for voltage is approached in a junction having only statistical variations in field, like those discussed in subsection III B.

Let us assume, in keeping with equation (2.32), that there is a region of area A in which V_B has uniformly its lowest breakdown value $V_{B\,\text{extreme}}$.

$$V_{B\,\text{extreme}} \simeq V_B - 5\delta V_B \qquad (5.1)$$

* Since this paper was prepared, the publications of KIKUCHI and TACHIHAWA[36] and KUKUCHI[37] have come to the author's attention. The importance of oxygen on microplasmas in silicon and germanium is discussed in these articles as the basis of a somewhat different model from that of subsection VC.

where V_B is the average breakdown voltage and

$$5\delta V_B \simeq (\ln N)^{1/2}\delta V_B \qquad (5.2)$$

where, from equation (3.23),

$$\delta V_B = C^{1/2}(qF_B/K)^{1/2}/2 \simeq C^{1/2}0{\cdot}14 \text{ volt} \qquad (5.3)$$

where δV_B is the root mean square deviation in voltage from cube to cube for the N cubes of the junction, and C is the compensation factor of equation (3.15).

If the applied voltage is greater than $V_{B\,\text{extreme}}$, then if one electron crosses the area, it will probably produce a larger second generation, and the successive generations of carriers will build up in a geometric series. The factor per generation is approximately S of equation (4.3). The space charge of these carriers will reduce the field, and a stable situation will occur when S is reduced to unity.

Fig. 16 represents the situation under the stable condition. In keeping with the value 0·32 for m in equation (4.1), we suppose the multiplication occurs chiefly in the middle third of the space-charge layer. Carriers move through this layer and through most of the junction with the limiting drift velocity given by solving equation (2.21) for v_d.

$$v_d = (E_R/m^*)^{1/2} \simeq 2\times 10^7 \text{ cm/sec} \qquad (5.4)$$

The field in the center one-third of the junction is reduced by the dipole layer due to unbalanced mobile charges in the outer thirds. Each charge is of magnitude

$$Q = IW_B/3v_d \qquad (5.5)$$

where $W_B/3v_d$ is the transit time and I is the total current. This sets up an electric field

$$F_I = Q/KA = IW_B/3Kv_dA \qquad (5.6)$$

In order to sustain this current, the applied voltage must be increased above $V_{B\,\text{extreme}}$, which gives $S = 1$, by an amount ΔV at the edges of the space-charge layer so that the field due to the moving carriers is neutralized:

$$\Delta V/W_B = F_I \qquad (5.7)$$

This leads to an effective series resistance R_{sc} due to space-charge

$$R_{sc} \equiv \Delta V/I = W_BF_I/I$$
$$= (W_B^2/A)(1/3Kv_d) \simeq$$
$$\simeq (W_B/A)\ 16{,}000 \text{ ohms} \qquad (5.8)$$

The quantity Kv_d is a characteristic space-charge conductance. Its value is approximately

$$Kv_d = 10^{-12} \text{ farad/cm} \times 2\times 10^7 \text{ cm/sec}$$
$$= 2\times 10^{-5} \text{ mho} \qquad (5.9)$$

If the area A has dimensions smaller than W_B, then a different limiting form must be used depending on the geometry. R_{sc} then increases less rapidly than $1/A$.

It should be noted that the resistance R_{sc} is an order of magnitude more important than the ohmic spreading resistance outside the space-charge layer. This latter can be estimated by considering that the current arrives at the edge of the space-charge layer over a circle of radius $r = (A/\pi)^{1/2}$

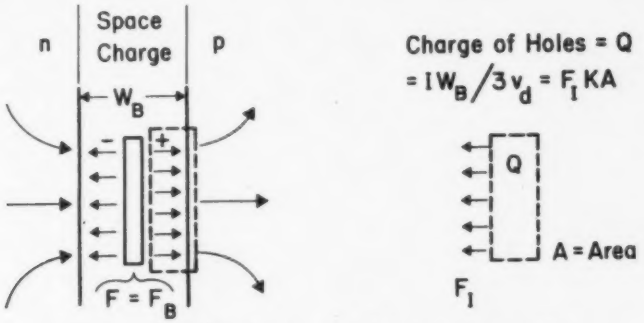


FIG. 16. Method of estimating the space-charge series resistance R_{sc}.

and is carried by material of varying conductivity, starting with values of the form $q\mu_p N_B$ at the edge of the space-charge layer, and having an effective average value of about twice this much. We represent the two sides by an average material of conductivity $2q\mu N_B$ where N_B is the charge density of equation (3.44) at the edge of the space-charge layer and $\mu = 2\mu_n\mu_p/(\mu_n+\mu_p)$ is a "series" average mobility. This leads to a spreading resistance of

$$R_{sr} = 2(4r2q\mu N_B)^{-1} \tag{5.10}$$

The 2 in the numerator represents the two sides of the space-charge layer. Using equation (3.44) to eliminate N_B gives

$$R_{sr} = [(\pi^{1/2}/16)(W_B^2/A)^{1/2}]/\mu F_B K \simeq$$
$$\simeq 400(W_B^2/A)^{1/2} \text{ ohms} \tag{5.11}$$

This resistance is much smaller than R_{sc} due to the term $\mu F_B K$ which occurs instead of $v_d K$. The velocity-like quantity $\mu F_B = 2{\cdot}5\times10^8$ cm/sec is more than one order of magnitude larger than v_d.

A characteristic current value for turned-on regions in the statistical junction is obtained by combining R_{sc} and δV_B. This gives

$$I = \delta V_B/R_{sc} = C^{1/2}(A/W_B^2)\,10^{-5} \text{ amp} \tag{5.12}$$

Thus, if the lowest extreme cube has a breakdown voltage $5\delta V_B$ below the average, it will carry a current of about 50 $C^{1/2}$ μamp when the voltage is raised to V_B, the average breakdown voltage.

As the voltage is raised above the extreme minimum, progressively increasing numbers of areas should reach the sustaining condition, in accordance with the distribution of equation (3.34). These should be stabilized by a high series resistance. The current should increase according to the same law as the number of areas, and thus for junctions of area about 10^{-3} cm^2 or more

$$I = 10^{-5}C^{1/2}\exp\{30C^{-1/2}[V - V_{B\,\text{extreme}}]\} \text{ amp} \tag{5.13}$$

This is approximately the behavior reported by SENITZKY and MOLL[33] and shown in Fig. 17. The slope is about twice as large as can be explained from equation (5.13). However, it is not clear whether their current was of the type described here or was due to a varying duty cycle for several "on–off" microplasmas. In the range where (5.13) applies, differential resistance times current should be constant or equal to $(C^{1/2}/30)$ volt.

Equation (5.13) should apply until V increases by about 2 or 3 times δV_B above $V_{B\,\text{extreme}}$. Above this point the exponential approximation to

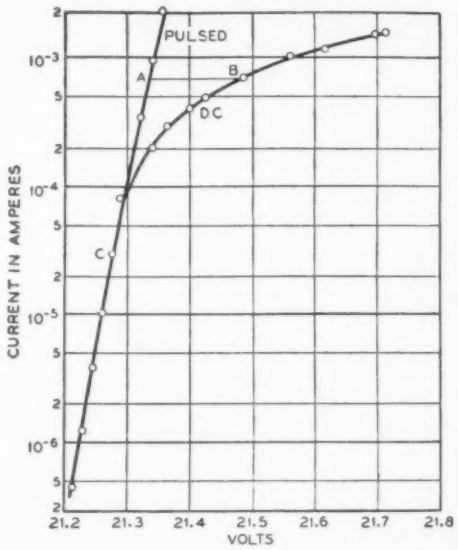


FIG. 17. Dependence of reverse current upon reverse voltage in a p–n junction near breakdown (SENITZKY and MOLL[33]). (The d.c. curve is thought to be affected by heating and can be disregarded.)

the Poisson distribution should be inexact. For voltages a few times δV_B higher than V_B, almost all of the area will be above its local breakdown. The differential conductance per unit area will then be given by equation (5.8) with A representing the entire junction area. Expressed as mho/cm^2, this becomes

$$dI/A\,dV = 3Kv_d/W_B^2 \simeq 7\times10^4/V_B^2 \text{ mho/cm}^2 \tag{5.14}$$

For a junction of 20-V breakdown and an area of 10^{-2} cm^2, this gives 2 mho or 0·5 Ω. The voltage which will produce this condition will be about $2\,\delta V_B$ above the breakdown V_B and will give a current of about $2\,\delta V_B$ times the differential conductance, or a current density of about

$$I/A = 3C^{1/2}(qF_B/K)^{1/2}Kv_d/W_B^2 \simeq$$
$$\simeq C^{1/2}3\times10^3/V_B^2 \text{ amp/cm}^2 \tag{5.15}$$

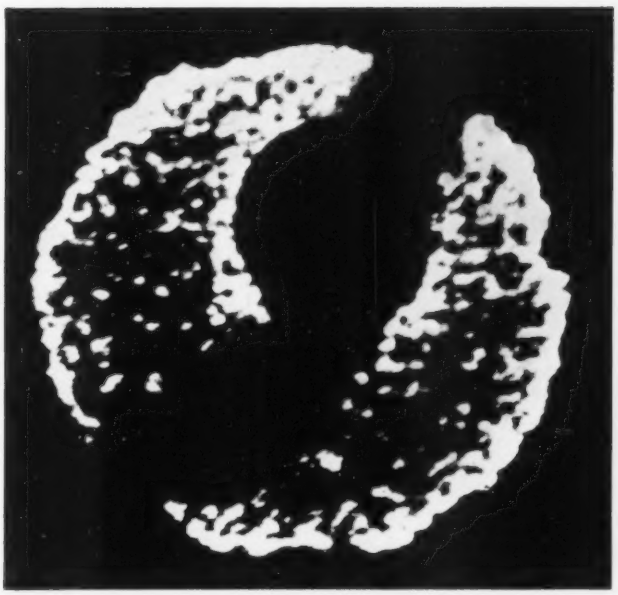


FIG. 18. Distribution of light from a uniform junction with a guard ring as observed by BATDORF *et al.*[28]

[*facing p.* 60

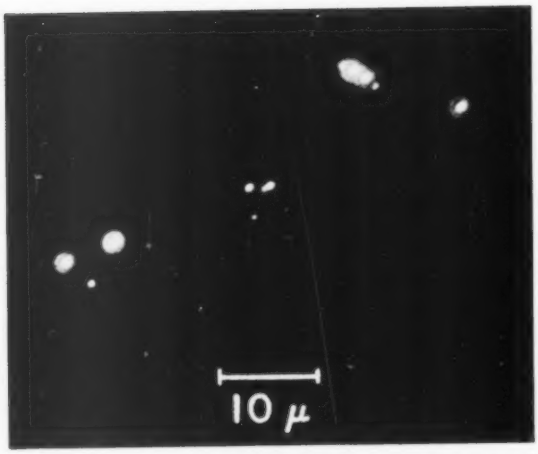


FIG. 22. Pattern of microplasma light observed by GOETZBERGER[42].

or about 0·6 A for the example considered with $C = 1$. Currents of about this order of magnitude should be necessary to override the statistical fluctuations and produce substantially uniform current density and uniform light emission.

Statistical turn-on of the type discussed here has apparently been observed by BATDORF *et al.*[28] using junctions having guard rings to prevent surface breakdown. Fig. 18 shows one pattern of light emission. No microplasma noise was observed from such a junction.

BATDORF *et al.* also show one junction which glowed uniformly. This is not to be expected since the statistics should prevent it. Similar uniform glows had previously been reported by MCDONALD *et al.*[30], also using guard rings. Subsequent investigation by GOETZBERGER[38] has shown that these uniform glows were due to injection from the back junction of an n–p–n structure with transistor-like α-values of perhaps 0·1. For such a structure, M values only as high as 10 are required, and for these V may be sufficiently below $V_{B\,\text{extreme}}$ that no area is so near its breakdown that it carries an exceptionally large current and thus appears as a glowing spot.

It is evident that more experiments involving, for example, making transistor-like structures and injecting carriers will be needed before the statistical model can be adequately tested.

V C. *The* SiO_2 *precipitate model for microplasmas*

We shall now consider the possibility that microplasma effects in silicon are due to precipitates of SiO_2. The requirements for a microplasma-forming imperfection were discussed at the end of subsection V A. These appear to be met by precipitated inclusions of SiO_2 as studied by KAISER[39].

KAISER reports that silicon heat treated at 1000°C contains precipitates of SiO_2, the density being about 5×10^{10} precipitates per cm^3 and the diameter being about 10^{-5} cm. At this density, a typical space-charge layer with area of 10^{-3} cm^2 and thickness 10^{-4} cm would contain 5000 precipitates. Specimens containing less oxygen and subject to different heat treatments might contain much smaller numbers.

If we take the precipitate to be a sphere, then, since its dielectric constant is much smaller than silicon, the disturbance in electric field will cause the field to be about 1·5 stronger around the equator, as represented in Fig. 19, than at several diameters away from the sphere. At an average field of $F_B = 500{,}000$ V/cm, the field near the sphere would thus be about 750,000 V/cm and the voltage drop in one sphere radius would be about 3 or 4 V. At such high fields, values of α as high as 10^5 cm^{-1} may be expected, so that αW values greater than unity [see equation (4.1)] might occur and a condition of sustained breakdown be possible. If no other effects occurred, this sustained breakdown should be stable and associated with a series resistance like that discussed in the previous subsection. It should not produce the "on-off" noise associated with microplasmas.

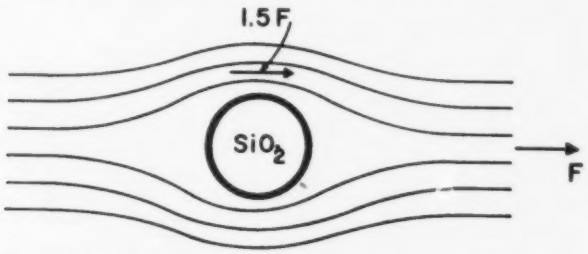


FIG. 19. The influence of a low dielectric constant SiO_2 sphere on increasing the electric field in the space-charge region of a p–n junction.

A "lock–on" mechanism may, however, be provided by the surface states at the precipitate silicon interface. These states are sufficient in number to produce electric fields of the order of the breakdown field. In order for a charge in the surface states to cause "lock–on" of the microplasma, it is necessary that these states become

charged so as to increase the strength of the electric field at the location of the highest field where carriers are being generated.

Fig. 20 illustrates this requirement in general terms. It represents a high-field region and two regions of traps. The charge in the traps produced by the carriers moving by them should be of

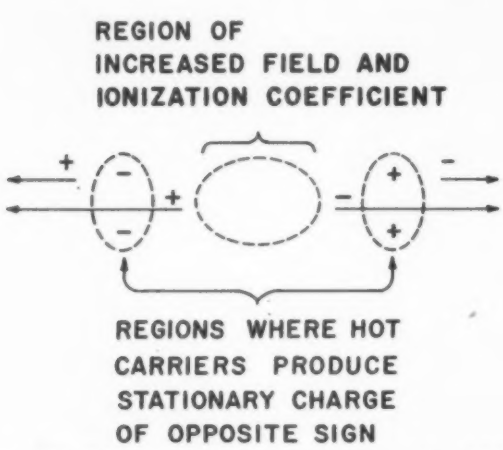


FIG. 20. Schematic representation of the way in which traps may enhance the field to produce microplasma "lock-on".

opposite sign from the carriers themselves so that the trapped charge will increase the field in the breakdown region and produce "lock–on". Can surface states play such a role?

It seems probable that surface states can acquire the needed reversed charges in the presence of hot carriers. The microplasma temperatures discussed in subsection II F correspond to energies of about 0·2 eV, and these are larger than the separation of about 0·1 to 0·15 eV of surface states from the band edges determined by STATZ *et al.*[40] Thus hot-carrier energies are quite high enough to ionize the surface traps. Depending upon the cross-section of the traps and their numbers, the charge produced might be either the same as or opposite to the hot carriers. If the sign of charge induced is opposite to the carriers, then "lock-on" can occur.

It is, of course, not necessary both that hot holes induce a negative charge on the traps and that hot electrons induce a positive charge. Either case singly can suffice to cause microplasma effects, since in general there will be unsymmetrical conditions due to the distribution of several precipitates or the shape of a single precipitate.

Fig. 21 illustrates an unsymmetrical example in which it is assumed that several precipitates have formed along a dislocation. It is assumed that hot electrons cause the small precipitate to become positively charged. It is also conceivable that the charged surface states which produce lock-on are on the same particle and are charged by secondaries generated beyond it.

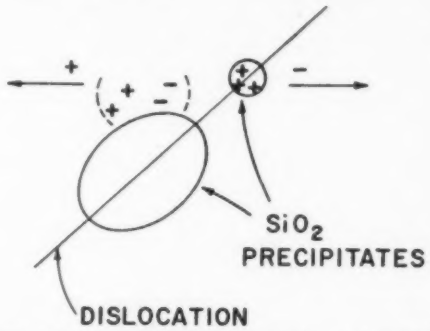


FIG. 21. Illustration of the co-operative action of two precipitates along a dislocation to produce "lock-on".

The type of mechanism proposed in Fig. 21 will be recognized as similar to that which occurs in vacuum electronics when secondary-electron emission plays a role. A secondary emitting surface exposed to electron bombardment will tend to become positively charged and may "run-away" towards the anode potential unless it is connected to a voltage source by sufficiently low impedance.

For the small dimensions, the order of 10^{-5} cm, the change in voltage associated with lock-on would only be of the order of a few volts. The step in current would, in accordance with equation (5.8), be of the order of 10^{-5} A. These are only in rough qualitative agreement with the experimental values quoted in subsection V A.

It is evident that a wide variety of models can be constructed along the lines just discussed. Since they are obviously of a highly speculative nature, we shall here only briefly indicate certain directions of investigation.

Effects similar to those discussed above may occur on the silicon–SiO_2 interface where the junction meets the surface. This may account for

the prevalence of microplasma formation on the surface.

Groups of small interacting precipitates may produce complex local patterns of breakdown. These may be what are observed by GOETZBERGER[41] and illustrated in Fig. 22.

Current densities in glowing microplasma spots observed by GOETZBERGER are of the order of 100 μA/μ^2 or 10^4 A/cm^2. This is of the same order as the current density expected to be associated with uniform glow from a statistical junction—see equation (5.15). The spreading of the light around the bright central spot shown in GOETZBERGER's paper appears to be real and not an optical artifact. Diffusion outwards of successive generations of secondaries does not appear large enough to explain it, and the structure is somewhat mysterious at present.

Mechanical damage, producing dense arrays of dislocations, may produce areas of sufficient density of traps to produce the same effects as just discussed for surface states.

In connection with dislocations, we shall examine the proposal made by CHYNOWETH and PEARSON[42] that the deformation potential of about 0·2 eV near a dislocation can have an important influence on microplasma formation. Owing to the wave nature of the electron, its binding energy to such a line of lowered potential will be much less than 0·2 eV. Since electron temperature from microplasma radiation is nearly as big as 0·2 eV, the deformation should not influence the motion of carriers in the space-charge region appreciably. (In neutral regions, dislocations can affect motions significantly, as shown by LOGAN *et al.*[43])

Perhaps the best way to close this section is to say that interesting phenomena can and do take place in the high and variable fields in reverse-biased p–n junctions. Many of the details of these do not yet have an established explanation in terms of atomic models.

VI. THE EFFECT OF PRECIPITATED METAL PARTICLES

Metal precipitates, since they are conducting, should behave very differently from the SiO_2 precipitates of subsection V C.

The work of GOETZBERGER and SHOCKLEY[29,44] has shown that the precipitation of metal particles in silicon results in a characteristic current–voltage curve. For such a case, the current increases as approximately the fifth or sixth power of the voltage over several decades of current, as shown in Fig. 23. The curve extends to voltages below 1 V. For this reason it is believed the effect does

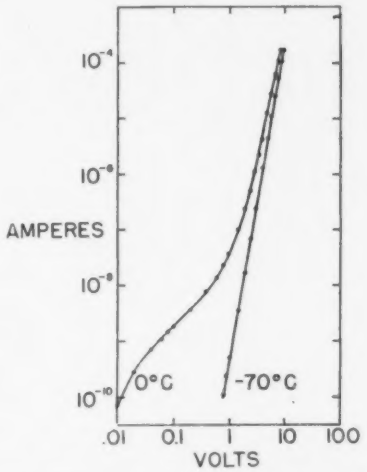


FIG. 23. Current–voltage characteristic of a soft junction containing precipitated copper (GOETZBERGER and SHOCKLEY[29]).

not have to do with avalanche which, if it were appreciable at such a low voltage, would lead to infinite multiplication at a slightly higher voltage. Instead it is conjectured that the mechanism of generating the current is tunnelling or Zener effect.

Evidence that the breakdown occurs at a localized spot in the p–n junction, probably corresponding to one precipitate, is obtained by using a potential-probing technique. This technique makes use of methods previously used, but on a finer scale, both so far as distances and voltages are concerned. Fig. 24 shows a potential plot for the junction measured in Fig. 23, at a current of 200 μA. This current enters the top diffused layer of the junction through one probe. Another probe is used as a voltage reference, and a third probe, not shown, was used to map the 10-μA contour lines. The lines of current flow are perpendicular to the equipotential lines, and

it is seen from the potentials that they must diverge from the current source and converge at an interior point in the specimen. The sheet conductance of the top diffused layer was approximately 1 mho per unit square. From this, it was possible to state that within experimental accuracy the current converging upon the localized sink was practically equal to the 200 μA flowing. Fig. 25 shows a plot for a similar junction.

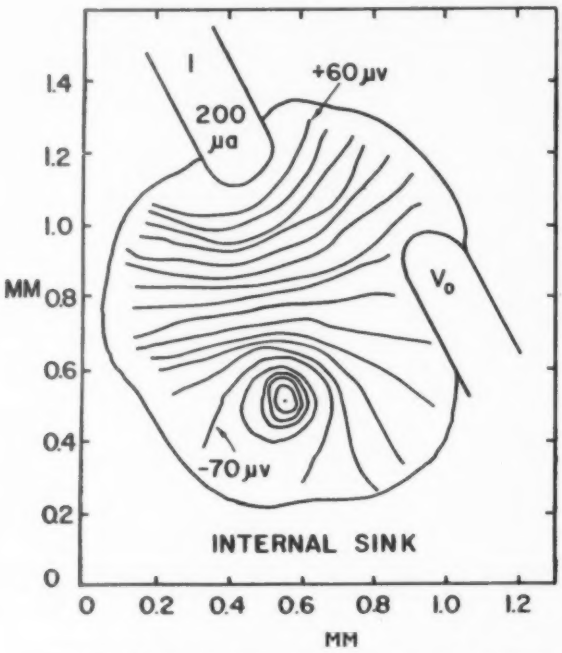


FIG. 24. Plot of equipotentials due to current flow in the thin diffused layer of the diode corresponding to Fig. 23.

In passing, it should be noted that the same potential-probing technique can be used for locating surface leakages when these exist, and two other cases of potential probing are shown in Figs. 26 and 27.

Several mysteries remain in connection with the current flow associated with precipitated metal particles. The fields in the neighborhood of the particle must be high enough to produce tunnelling, and this is higher than is necessary to produce avalanche multiplication. Yet no significant avalanche multiplication effect appears to arise, for, if it did, it would be expected that the current would approach infinity and be limited by a series resistance at some voltage substantially lower than the general breakdown voltage.

The considerations of ohmic resistance due to space charge should apply in such cases, and it is difficult to see why ohmic resistances of the order of tens of thousands of ohms or more are not associated with the observed effect. The limiting ohmic resistance was not measured for the specimen of Figs. 23 and 24 since the resulting heating of similar specimens caused a destruction of the phenomenon. It seems probable, however, that a series resistance of as large as 10,000 Ω would have been appreciable in connection with Fig. 23.

It is evident in the case of these effects that further theoretical work will be necessary to see whether or not a reasonable model of a metal precipitate can give rise to the observed current–voltage relationship.

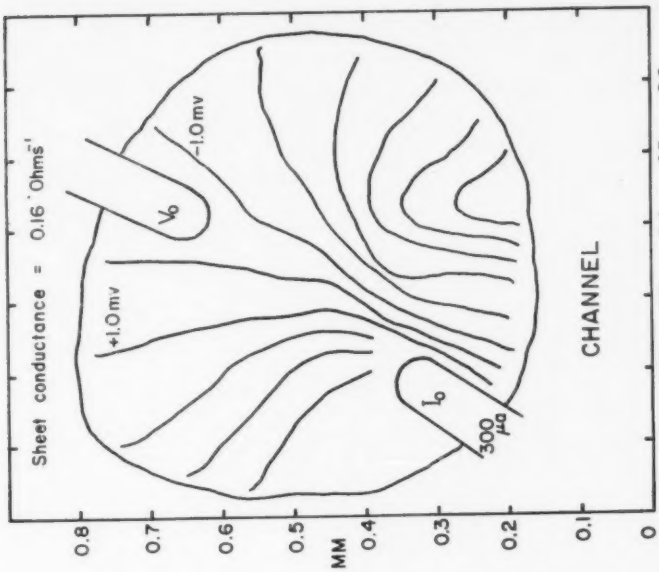


FIG. 26. An illustration of the potential-probing method for locating a current sink on the surface of a p–n junction.

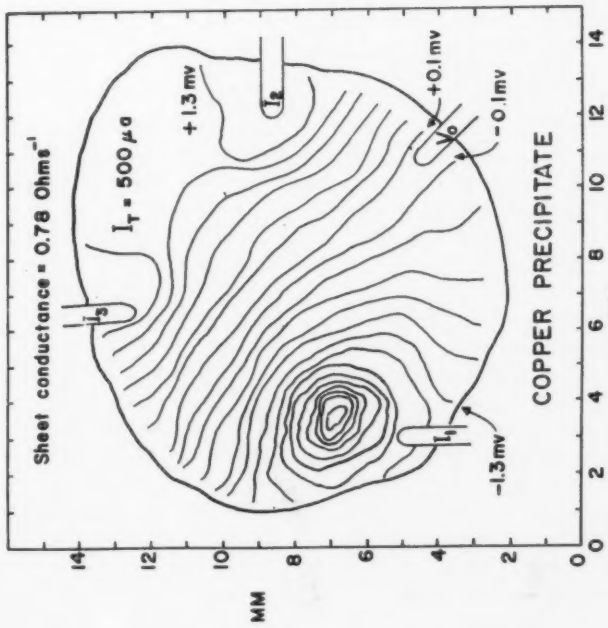


FIG. 25. Convergence of current flow on a precipitated copper precipitate in a larger-area p–n junction.

It should be remarked, in passing, that the effect of metal precipitates in producing excess or soft localized currents in p–n junctions is of commercial significance. The work of GOETZBERGER[45] and GOETZBERGER and SHOCKLEY[44] has shown that these precipitates may actually be removed from a junction in which they are formed, by subsequent heat treatment with a glassy phase upon the surface.

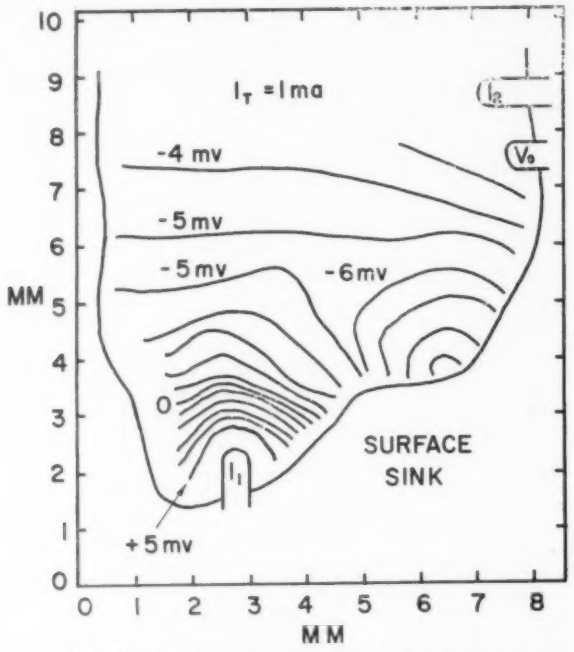


FIG. 27. Another example of a surface sink on a p–n junction.

VII. CONCLUSIONS

A simple model of the behavior of holes and electrons in the high reverse fields of p–n junctions has been shown to explain a large body of experimental information. How such a simple model can work so well is somewhat of a mystery in view of the known complications of the band structure.

Statistical fluctuations of donors and acceptors are shown to have significant effects on the behavior of p–n junctions in the region of avalanche breakdown. Some of these are calculated, and areas for investigation are suggested.

It is concluded that microplasma effects, including the "lock-on" or negative-resistance phenomena and noise, require structural imperfections of dimensions of the order of a few hundred to a thousand Angstroms, together with traps capable of storing and immobilizing charge. SiO_2 precipitated inclusions are shown to have the necessary attributes, and it is suggested that they are a possible cause of microplasma effects.

Metal precipitates are shown to behave quite differently, and some problems and answers related to their behavior are discussed.

The problems dealt with are of importance, both from the point of view of their scientific interest and because they bear on the problems of producing useful semiconductor devices. It is to be hoped that their study will be a contribution of the sort the President of the Czechoslovakian Academy of Sciences described as being of benefit to both science and society.[46]

Acknowledgements—Figs. 6, 13(e) and 18 were reproduced from the *Journal of Applied Physics* by courtesy of

The American Institute of Physics; Figs. 8, 13(a), 14 and 17 were reproduced from *Physical Review* by courtesy of the Editors; the data of Fig. 13(d) were reproduced by courtesy of the American Physical Society; and Figs. 13(b) and (c) were reproduced from *Solid State Physics and Electronics Telecommunications* (Vol. 1) by courtesy of the Academic Press.

The author has been helped while preparing this article by stimulating discussions, critical comments and assistance in preparing and checking the figures and the references. It is a pleasure to express appreciation to A. G. CHYNOWETH, G. C. DACEY, W. FRANZ, A. GOETZBERGER, R. R. HENLEY, J. L. MOLL, J. SIMONS, C. L. STEPHENS and B. M. WUL.

REFERENCES

1. W. SHOCKLEY, *Bell Syst. Tech. J.* **30,** 990 (1951).
2. P. A. WOLFF, *Phys. Rev.* **95,** 1415 (1954).
3. H. PALEVSKY, D. J. HUGHES, W. KLEY and E. TUNKELO, *Phys. Rev. Letters* **2,** 258 (1959).
4. B. N. BROCKHOUSE and P. K. IYENGAR, *Phys. Rev.* **111,** 747 (1958).
5. V. S. VAVILOV, *J. Phys. Chem. Solids* **8,** 223 (1959).
6. J. TAUC, *J. Phys. Chem. Solids* **8,** 219 (1959).
7. S. KOC, *Czech. J. Phys.* **6,** 668 (1956).
8. S. KOC, *Czech. J. Phys.* **7,** 91 (1957).
9. J. TAUC and A. ABRAHÁM, *Czech. J. Phys.* **9,** 95 (1959).
10. E. ANTONČÍK, *Czech. J. Phys.* **7,** 674 (1957).
11. E. ANTONČÍK, *Czech. J. Phys.* **8,** 492 (1958).
12. K. G. MCKAY and K. B. MCAFEE, *Phys. Rev.* **91,** 1079 (1953).
13. A. G. CHYNOWETH, *J. Appl. Phys.* **31,** 1161 (1960).
13a. A. G. CHYNOWETH, *Phys. Rev.* **109,** 1537 (1958).
14. A. G. CHYNOWETH and K. G. MCKAY, *Phys. Rev.* **108,** 29 (1957).
15. L. V. KELDȲSH, V. S. VAVILOV and K. I. BRITSIN, *International Conference on Semiconductor Physics, Prague, August,* 1960, P/J-16.
16. W. FRANZ, *Z. Naturf.* **13a,** 484 (1958).
17. R. J. GRAINGER, J. W. MAYER, J. S. WIGGINS and S. S. FRIEDLAND, *Bull. Amer. Phys. Soc.* **5,** 265 (1960).
18. A. ZAREBA, *Bull. Acad. Pol. Sci.* **7,** 255 (1959).
19. A. ZAREBA. Personal communication (1960).
20. A. G. CHYNOWETH and K. G. MCKAY, *Phys. Rev.* **102,** 369 (1956).
21. J. A. BURTON, *Phys. Rev.* **108,** 1342 (1957).
22. B. SENITZKY, *Phys. Rev.* **116,** 874 (1959).
23. W. SHOCKLEY, *Bull. Amer. Phys. Soc.* **5,** 161 ((1960).
24. E. J. GUMBEL, *Statistics of Extremes.* Columbia University Press, New York (1958).
25. K. G. MCKAY, *Phys. Rev.* **94,** 877 (1954).
26. S. L. MILLER, *Phys. Rev.* **105,** 1246 (1957).
27. B. M. WUL and A. P. SHOTOV, *Solid State Physics and Electronics Telecommunications* Vol. 1, p. 491. Academic Press, New York (1960).
28. R. L. BATDORF, A. G. CHYNOWETH, G. C. DACEY and P. W. FOY, *J. Appl. Phys.* **31,** 1153 (1960).
29. A. GOETZBERGER and W. SHOCKLEY, *Structure and Properties of Thin Films* p. 298 (Ed. by C. A. NEUGEBAUER *et al.*). John Wiley, New York (1959).
30. B. MCDONALD, A. GOETZBERGER and C. STEPHENS, *Bull. Amer. Phys. Soc.* **4,** 455 (1959).
31. A. G. CHYNOWETH and K. G. MCKAY, *J. Appl. Phys.* **30,** 1811 (1959).
32. K. S. CHAMPLIN, *J. Appl. Phys.* **30,** 1039 (1959).
33. B. SENITZKY and J. L. MOLL, *Phys. Rev.* **110,** 612 (1958).
34. D. J. ROSE, *Phys. Rev.* **105,** 413 (1957).
35. A. F. GIBSON and J. R. MORGAN, *Solid-State Electron.* **1,** 54 (1960).
36. M. KIKUCHI and K. TACHIHAWA, *J. Phys. Soc. Japan* **14,** 1830 (1959); **15,** 835 (1960).
37. M. KIKUCHI, *J. Phys. Soc. Japan* **15,** 1822 (1960).
38. A. GOETZBERGER, *J. Appl. Phys.* In press (1960).
39. W. KAISER, *Phys. Rev.* **105,** 1751 (1957).
40. H. STATZ, G. A. DEMARS, L. DAVIS, JR., and A. ADAMS, JR., *Semiconductor Surface Physics* (Ed. by R. H. KINGSTON) pp. 139–168. University of Pennsylvania Press (1957).
41. A. GOETZBERGER, *International Conference on Semiconductor Physics, Prague, August,* 1960, P/Q-4.
42. A. G. CHYNOWETH and G. L. PEARSON, *J. Appl. Phys.* **29,** 1103 (1958).
43. R. A. LOGAN, G. L. PEARSON and D. A. KLEINMAN, *J. Appl. Phys.* **30,** 885 (1959).
44. A. GOETZBERGER and W. SHOCKLEY, *J. Appl. Phys.* **31,** 1821 (1960).
45. A. GOETZBERGER, *Bull. Amer. Phys. Soc.* **5,** 160 (1960).
46. Z. NEJEDLÝ, *International Conference on Semiconductor Physics, Prague, August,* 1960.

Solid-State Electronics Pergamon Press 1961. Vol. 2, No. 1, pp. 68–76. Printed in Great Britain

NOTES

Field-effect pinch-off of surface leakage in high-voltage diodes

(*Received* 18 *May* 1960; *in revised form* 25 *July* 1960)

A LARGE-AREA and high-back-voltage germanium diode, which is not affected by exposing the p–n junction to the atmosphere, has been developed with a particular geometrical construction. It is well known that the reverse characteristics in the conventional diode are so greatly affected by the surface-leakage current that its breakdown voltage is much lower than the value theoretically expected.

Because the depletion layer has an insulating property, it could be possible to reduce the surface leakage by pinching off this current path with the depletion layer. In order to investigate this pinch-off effect, we have fabricated a diode with 20 Ω-cm n-type germanium, whose construction is shown in Fig. 1. The periphery of the p–n junction was

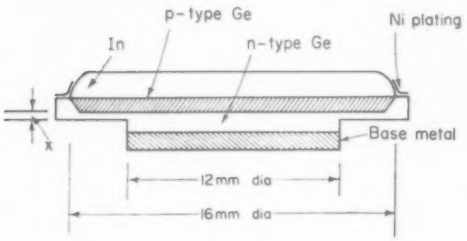


FIG. 1.

plated with nickel to produce a poor-rectifying contact, while the bottom surface of the germanium wafer was chemically etched in parallel with the p–n-junction surface. The plated nickel surface was coated with wax to prevent its being etched. The reverse characteristics of the diodes, measured as a function of the thickness x of the n-type germanium layer indicated in Fig. 1, are shown in Fig. 2.

One can clearly see that both the saturation current and the pinch-off voltage are smaller in the

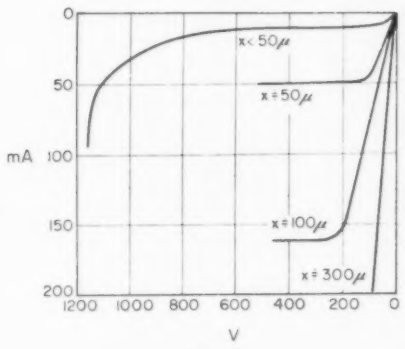


FIG. 2.

unit of the narrow layer. In the case of $x < 50\ \mu$, the poor-rectifying nickel plating has no effect upon the performance of the diode, and the breakdown voltage has increased to more than 1000 V, although the breakdown voltage of the conventional construction diode made of the same material is only from 300 to 500 V. In the case of $x > 50\ \mu$, both the saturation current and the pinch-off voltage are fairly large. On the other hand, for the conventionally constructed diode, the saturation current is usually as much as 1·0 mA/cm² and the saturation voltage is only about 0·025 V.

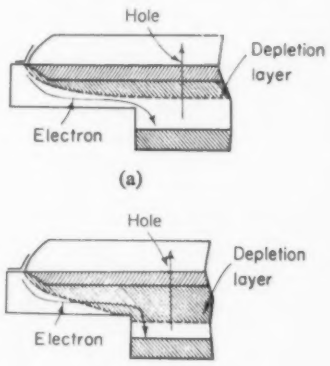


FIG. 3. (a) At low bias. (b) At high bias.

FIG. 3. Mesas obtained. One division is 100μ. The small one is about 30μ in diameter.

From these results, it could be considered that the surface-leakage current flows as shown in Fig. 3(a), because the depletion layer is narrow, so long as the applied voltage is low. When a high voltage is applied, the leakage current is pinched off by the depletion layer extending to the edge of the layer, as illustrated in Fig. 3(b).

In conclusion, one can achieve reduced surface leakage and increased reverse-breakdown voltage by means of proper geometry so as to utilize the field-effect pinch-off.

T. KAN

Shindengen Research Division
Hanno, Saitama, Japan

Micromasking for chemical etching of semiconductors

(*Received* 22 *July* 1960)

AN ETCHING technique called "mesa etch" has great significance for fabricating diffused-base high-frequency transistors and diodes for microwave and parametric use.

Mesas down to 30μ or less in diameter can be obtained by a method described here without using the troublesome photomasking technique which is usually employed. As shown in Fig. 1, a specimen of a semiconductor on which mesa etching should be done is placed on a platinum ribbon heater having a small thermal time constant,

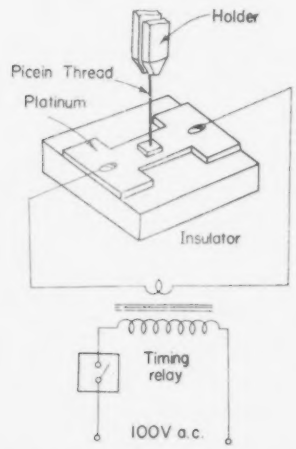


FIG. 1. A picein thread makes a micromask by pulsive firing.

i.e. the order of several tenths of a second. A fine thread made of picein, black wax used for sealing of vacuum systems, is put in touch on the predetermined spot with slight pressure by using a suitable mechanical holder under a microscope. The tip of the thread melts and makes a hemisphere when a current pulse is given to the heater, as shown in Fig. 2. Pulse length is determined by

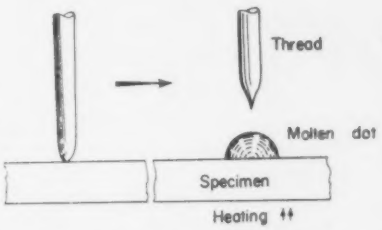


FIG. 2. A process of melting. Molten part is separated naturally from the thread.

a mechanical timing relay. Maximum temperature reached is a function of the current and the pulse length, ranging from 100 to 700 msec. Pulses longer than 1 sec spread the masking area too much.

Since the diameter of the masking dot, and hence that of the mesa, is about 1·5–3 times that of the thread according to experiment, very fine threads are required. Threads down to 20μ can be obtained by extending a piece of molten picein quickly with a small soldering iron.

The masked pellet is then etched in common etchant: CP–4 for germanium, or a mixture of HNO_3 and HF for silicon. The height of the mesa should be limited to one third of the diameter to avoid removing of the mask during etching. The picein mask can be dissolved easily by benzol after etching. Fig. 3 is a photograph of two mesas obtained by means of this technique, where one division of the scale is 100μ.

The effects of oxidation and contamination on the semiconductor surface may be neglected because of low temperature and short firing duration. This technique can be applied to other cases when circular masks are required.

Electrotechnical Laboratory S. DENDA
Nagata-cho, Chiyoda
Tokyo, Japan

Programming the zone-refining equation for a digital computer

(*Received* 22 *July* 1960; *in revised form* 12 *October* 1960)

The zone-refining equation

The distribution of an impurity (having a segregation coefficient k) after the nth zone pass is given by the equation[1]

$$\frac{d}{dq}C_n(q)+\frac{k}{l}C_n(q)=\frac{k}{l}C_{n-1}(q+l) \qquad (1)$$

where q is the distance along the rod, l is the zone length and $C_{n-1}(q)$ is the impurity distribution at q after the $(n-1)$th zone pass. Equation (1) may be written in the form

$$C_n(x)=ke^{-kx}\left[\int_0^1 C_{n-1}(p+1)e^{kp}dp+\int_0^1 C_{n-1}(p)dp\right] \qquad (2)$$

where $x=q/l$, and p is the integration variable.

This equation represents zone segregation for $N-1$ zone lengths, i.e. $0\leqslant x\leqslant N-1$. The impurity distribution in the last or Nth zone length, $N-1\leqslant x\leqslant N$, where directional freezing occurs, is given by

$$C_n(x)=C_n(N-1)(N-x)^{k-1} \qquad (3)$$

It may be seen from equation (3) that when $k<1$, the impurity distribution becomes asymptotic as x approaches N. Since it is required to perform a numerical integration in the region $N-1\leqslant x\leqslant N$ in order to obtain the profile in the region $N-2\leqslant x\leqslant N-1$ after the next zone pass, it is necessary to remove this singularity from the region of integration.

In this case, a process of "subtracting out" the singularity results in the following three general equations, which together define the segregation process resulting from zone refining.

$$C_n(x)=ke^{-kx}\left[\int_0^x C_{n-1}(p+1)e^{kp}dp+\int_0^1 C_{n-1}(p)dp\right] \qquad (0\leqslant x\leqslant N-2) \qquad (4)$$

$$C_n(x)=ke^{-kx}\left\{\frac{C_n(N-2)}{ke^{-k(N-2)}}+C_{n-1}(N-1)\left[\int_0^{x-N+2}\frac{e^{kt}-e^k}{(1-t)^{1-k}}dt+\frac{e^k}{k}\left(1-(N-1-x)^k\right)\right]\right\} \qquad (5)$$

$$(N-2\leqslant x\leqslant N-1)$$

$$C_n(x)=C_n(N-1)\,(N-x)^{k-1} \qquad (N-1\leqslant x\leqslant N) \qquad (6)$$

A programme has been written for the "Pegasus" digital computer, based on equations (4–6), which calculates impurity concentrations for any ratio of zone length to ingot length and for any value of segregation coefficient k. The starting

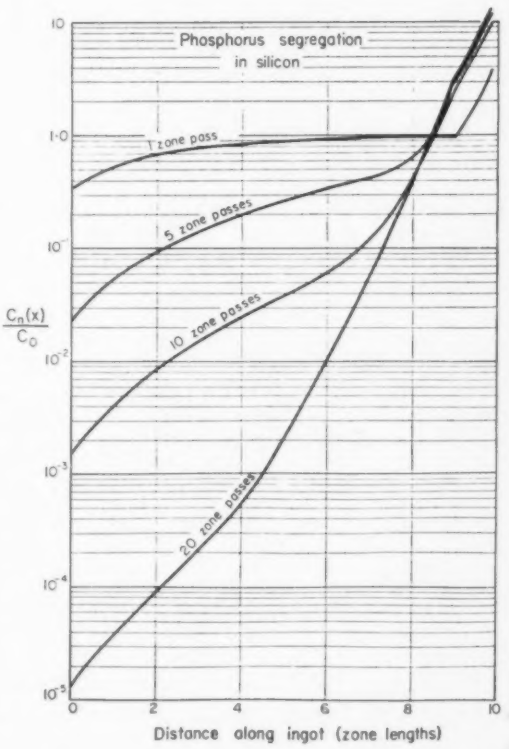


FIG. 1.

Table 1

Distance in zone lengths	$C_n(x)/C_0$ for no. of zone passes			
	1	5	10	20
0	0·350000	0·023026	0·001549	0·000013
0·25	0·404458	0·029458	0·002035	0·000017
0·50	0·454353	0·036635	0·002607	0·000022
0·75	0·500068	0·044560	0·003273	0·000029
1·00	0·541953	0·053230	0·004042	0·000036
2·00	0·677220	0·095099	0·008337	0·000087
3·00	0·772540	0·147349	0·015042	0·000201
4·00	0·839712	0·207835	0·024689	0·000544
5·00	0·887047	0·273983	0·038513	0·002071
6·00	0·920403	0·343246	0·060370	0·010654
7·00	0·943909	0·420316	0·119172	0·062455
8·00	0·960473	0·615804	0·418054	0·387234
9·00	0·972146	2·509969	3·047708	3·229884
9·875	3·756125	9·697879	11·775567	12·479449

condition is that the initial impurity distribution is uniform throughout the ingot, i.e. C_0.

Example

The impurity distributions in Table 1 for one, five, ten and twenty zone passes were computed for an impurity having a segregation coefficient $k = 0{\cdot}35$, e.g. phosphorus in silicon, and a ratio of zone length/ingot length (i.e. $1/N$) of 1 : 10.

These results are shown graphically in Fig. 1, from which it can be seen that after twenty zone passes the curves are tending towards the ultimate distribution.

Curves such as those in Fig. 1 have been computed for a number of impurity elements in silicon, for which the values for k are accurately known. By compounding two or more of these curves, theoretical profiles of electrical conductivity (a function of nett impurity content) have been obtained. Such curves are invaluable for analytical experiments where impurity concentrations in a given sample of material are determined from electrical resistivity measurements after a zone-refining operation.[2]

Acknowledgements—The authors are indebted to their colleagues at AEI, in particular Mr. S. MARSHALL of the Research Laboratory (Rugby) and Mr. G. DUCKWORTH of the Electronics Division (Leicester), for their aid and helpful discussions during the preparation of this paper.

Research Laboratory J. L. PARMEE
AEI (Rugby) Ltd.
Engineering Maths. Dept. J. A. ROBERSON
AEI (Rugby) Ltd.

References

1. N. LORD, *J. Metals, N.Y.* **5**, 1531, (1953).
2. J. L. PARMEE, *Proc. Rugby Engng. Soc.* **50**, 41 (1960)

Some effects of material parameters on the design of surface space-charge varactors

Introduction

It has long been realized[1,2] that the capacitance between a metal and a semiconductor varies with voltage owing to the behavior of the space-charge layer near the surface of the semiconductor.[3,4] Such capacitors would be useful for high-frequency parametric amplification provided that: (a) the fixed series capacitance can be made large enough so that the variation in overall capacitance is appreciable, and (b) the series resistance can be made small enough so that the cut-off frequency is high. For example, one solution to (a), suggested

by PFANN and GARRETT[5], is the use of a thin layer of oxide grown on the semiconductor as the dielectric spacer.

The present Note is a discussion of some of the effects of materials parameters (semiconductor and dielectric materials and layer thicknesses) on the capacitance variation and cut-off frequency.

Theory

1. *Surface space-charge layer.* The surface charge density Q_s and surface field strength E_s in the semiconductor at equilibrium are given[4] by

$$Q_s = \kappa_1\epsilon_0 E_s = \kappa_1\epsilon_0(kT/qL_D)F(u_s,u_B) \quad (1)$$

where

$$F(u_s,u_B) = \pm\sqrt{2}[(u_B-u_s)\sinh u_B+\cosh u_s-\cosh u_B]^{1/2} \quad (2)$$

(minus sign for $u_s>u_B$, plus sign for $u_s<u_B$), where $u = (E_f-E_i)/kT$; subscript B denotes bulk value, s surface value; E_f = Fermi energy, E_i = intrinsic Fermi energy, κ_1 = dielectric constant of the semiconductor and L_D = intrinsic screening length.

Equation (2) is based on the assumption that the semiconductor thickness $L_1 \gg$ the space-charge layer thickness. A suitable measure of the latter is the bulk screening length

$$L_B = L_D(\cosh u_B)^{-1/2} \quad (2a)$$

It is also assumed in equation (2) that the surface layer remains non-degenerate, but in the present context the difference is slight and will be ignored. For more exact calculations, the formulae of SEIWATZ and GREEN[6] may be used.

The capacitance per unit area of the space-charge layer is

$$C_1 = \frac{dQ_s}{dV_s} = -\frac{\kappa_1\epsilon_0}{L_D}\frac{\partial F}{\partial u_s} = -\frac{K_1\epsilon_0}{L_D}\frac{\sinh u_s-\sinh u_B}{F(u_s,u_B)} \quad (3)$$

Values of some of the parameters in the foregoing equations are given in Table 1.

Table 1. Values of parameters at 300°K

	L_D (cm)	kT/qL_D (V/cm)	$\kappa_1\epsilon_0/L_D$ (F/cm²)
Ge	7×10^{-5}	370	$2{\cdot}0\times10^{-8}$
Si	$2{\cdot}4\times10^{-3}$	12	$4{\cdot}5\times10^{-10}$
InSb	$2{\cdot}3\times10^{-6}$	$1{\cdot}1\times10^{4}$	6×10^{-7}

For $u_s = u_B$, C_1 becomes

$$C_{10} = \kappa_1\epsilon_0(\cosh u_B)^{1/2}/L_D = \kappa_1\epsilon_0/L_B \quad (4)$$

For n-type material with n electrons per cm³

$$C_{10} = q(\kappa_1\epsilon_0 n/kT)^{1/2}$$
$$7{\cdot}4\times10^{-16}(300\,\kappa_1 n/T)^{1/2}\ \text{F/cm}^2 \quad (4a)$$

The minimum value of C_1 is

$$C_{1\,\min} = \kappa_1\epsilon_0(\cosh u_m)^{1/2}/L_D \quad (5)$$

where u_m is the non-trivial solution of the transcendental equation

$$\cosh u_m = [(\sinh u_m-\sinh u_B)/F(u_m,u_B)]^2 \quad (6)$$

For $u_B \leqslant 30$, $C_{1\,\min}$ is fitted within a few per cent by

$$C_{10}/C_{1\,\min} = 1+0{\cdot}542u_B-0{\cdot}0115u_B^2+1{\cdot}25\times10^{-4}u_B^3 \quad (6a)$$

In parallel with the space-charge capacitance there is, almost inevitably, a capacitance due to surface states. However, at frequencies greater than about 50 Mc/s[7], this may be ignored since the states are unable to follow the alternations of the field. Their only effect will be to bias the surface, corresponding to the average trapped charge. Since the rate of trapping by empty states during one half-cycle will be much greater than the rate of release during the subsequent half-cycle, the bias will, in general, depend on the voltage amplitude and will require a few cycles to establish itself. During this time, as pointed out by PFANN and GARRETT[5], there will also be a net generation of minority carriers required for an inversion layer, so that the equilibrium expressions derived above will be reasonably accurate even near the minimum capacitance.

2. *Overall capacitance.* The dielectric layer of thickness L_2 and dielectric constant κ_2 will have capacitance per unit area, field strength and voltage drop given by

$$C_2 = \kappa_2\epsilon_0/L_2 \ , \ E_2 = \kappa_1 E_s/K_2 \tag{7}$$

$$V_2 = L_2E_2 = -(kT/q)(\kappa_1\epsilon_0/C_2L_D)F(u_s,u_B) \tag{7a}$$

where the constant corresponds to the average charge trapped in surface states, and will here be taken as zero. Fig. 1 illustrates the C–V curves for n-germanium.

3. *Cut-off frequency.* Owing to the finite conductivity of the semiconductor layer, there will be a cut-off frequency $\nu_m = (2\pi RC)^{-1}$, where

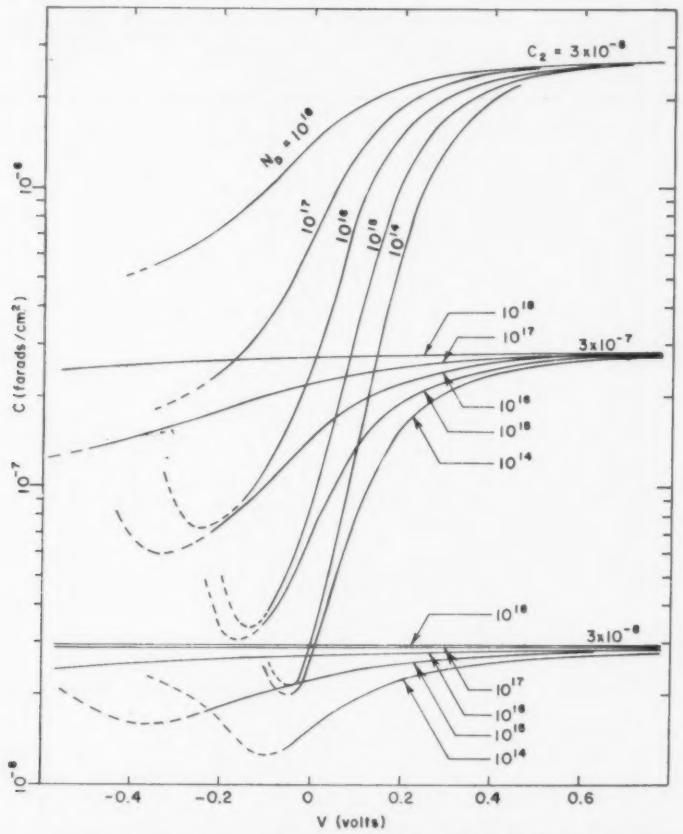


FIG. 1. Values of ν_m' [equation (10b)], L_B [equation (2a)] and C_{10} [equation (4)] for n-type germanium, silicon, and indium antimonide at 300°K.

Then the overall capacitance per unit area is

$$C(u_s,u_B) = C_2[1+C_2/C_1(u_s,u_B)]^{-1} \tag{8}$$

$$C_{\max}/C_{\min} = 1+C_2/C_{1\min} \tag{8a}$$

and the total voltage is

$$V = (kT/q)(u_s-u_B)+V_2+\text{ const.} \tag{9}$$

R = equivalent series resistance of unit area. The Q-factor (ratio of reactance to resistance) at any frequency ν is simply ν_m/ν. Since we assume $L_1 \gg L_B$, we may, except at extremely high frequencies of the order of the collision frequency, ignore the shunt resistance of the space-charge layer itself and take $R = \rho L_1$, where ρ = bulk

resistivity of the semiconductor $= (nq\mu)^{-1}$ (n = total carrier concentration, μ = weighted average mobility). Hence,

$$\nu_m = \nu_{m0} C_{10}(C_2+C_{10})/C_1(C_2+C_1) \quad (10)$$

where

$$\nu_{m0} = \nu_m'(L_B/L_1)(1+C_{10}/C_2) = \quad (10a)$$

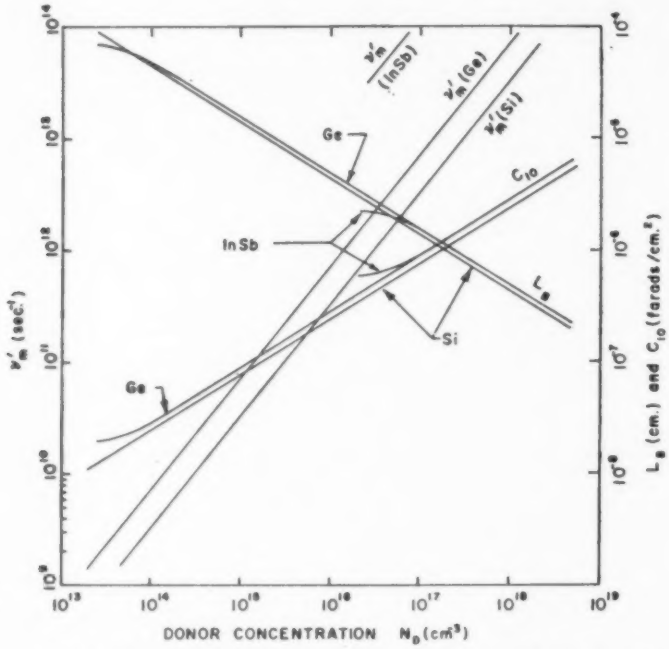


FIG. 2. Capacitance vs. voltage for n-germanium capacitors (donor concentration N_D). Dotted portions correspond to inversion layers.

zero-voltage cut-off frequency, and

$$\nu_m' = nq\mu/2\pi\kappa_1\epsilon_0 \quad (10b)$$

In Fig. 2, ν_m', L_B and C_{10} are plotted vs. carrier concentration for n-type germanium, silicon, and indium antimonide.

Discussion

Fig. 2 summarizes some of the potentialities, as well as the fabrication difficulties, of these devices. Extremely high-frequency operation appears possible if sufficiently thin layers of semiconductor and dielectric can be prepared. To illustrate, Fig. 3 presents values of ν_{m0} and C_{max}/C_{min} computed for germanium taking $L_1 = 10^{-4}$ cm, a thickness that is presently within reach by refined grinding or etching techniques. For other thicknesses, ν_m is inversely proportional to L_1, hence the use of thin evaporated layers of semiconductor would be decidedly advantageous. Other means of increasing the cut-off frequency without loss in capacitance ratio would be to use InSb (cf. Fig. 2) or to operate at a somewhat reduced temperature, since ν_m goes directly as the carrier mobility.

It is instructive to compare these devices with junction diodes. Since the equations governing the space-charge distribution are the same, a surface capacitor is analogous to one half of an abrupt junction diode. The essential difference is that the forward conduction is prevented by the dielectric layer, so that advantage can be taken of the large increase of C_1 under forward bias.

Probably one of the most formidable problems in the fabrication of these devices will be the achievement of the high C_2 required. The choice

of dielectric material to accomplish this will be governed by a variety of factors; these will not be discussed here except to point out that high dielectric constant materials, such as some of the titanates, appear advantageous in two respects; both L_2 and E_2 are inversely proportional to κ_2 for a given C_2. For example, for $C_2 = 10^{-6}$ F/cm^2, a film of $SrTiO_3$ ($\kappa_2 = 310$) would be about $0{\cdot}3\mu$ thick and would be subjected to a field of the order of 10^3 V/cm to drive C nearly to the maximum, whereas an oxide or ordinary organic layer would have to be as thin as about 40 Å and to withstand about 10^5 V/cm. However, other considerations, such as ease of preparation, might well override these.

The foregoing discussion has dealt with high-frequency performance. However, there may well be applications for these devices in the 100 Mc/s region, where very large capacitance ratios appear possible by using low carrier concentrations. Here, silicon would be advantageous. In making calculations for this range, the restriction $L_1 \gg L_B$ must be borne in mind, otherwise the formulae become much more complicated.

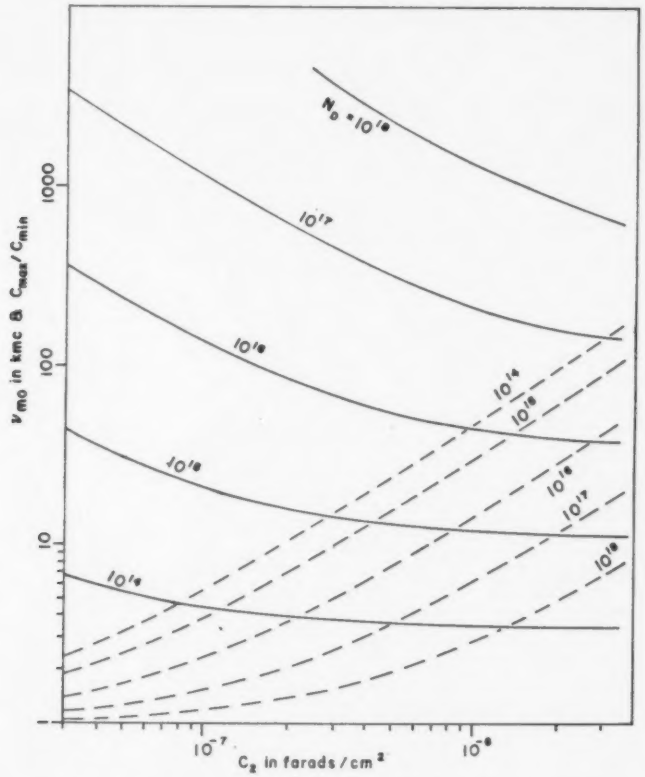


FIG. 3. Capacitance ratio C_{max}/C_{min} (broken lines) and zero-voltage cut-off frequency ν_{m0} (solid lines) for n-germanium capacitors.

Acknowledgements—Thanks are due to A. L. BROWN and Mrs. A. CALHOUN for assistance with the computations, and to R. I. HARRISON for a discussion of varactor performance specifications.

General Telephone & Electronics Laboratories, Inc.
Bayside
New York

D. R. FRANKL

References

1. W. L. Brown, W. H. Brattain, C. G. B. Garrett and H. C. Montgomery, *Semiconductor Surface Physics* (Edited by R. H. Kingston) p.117. University of Pennsylvania Press, Philadelphia (1956).
2. H. M. Bath and M. Cutler, *Bull. Amer. Phys. Soc.* II **3**, 378 (1958).
3. C. G. B. Garrett and W. H. Brattain, *Phys. Rev.* **99**, 376 (1955).
4. R. H. Kingston and S. F. Neustadter, *J. Appl. Phys.* **26**, 718 (1955).
5. W. G. Pfann and C. G. B. Garrett, *Proc. I.R.E.* **47**, 2011 (1959).
6. R. Seiwatz and M. Green, *J. Appl. Phys.* **29**, 1034 (1958).
7. H. C. Montgomery, *Phys. Rev.* **106**, 441 (1957).

Solid-State Electronics Pergamon Press 1961. Vol. 2, No. 1, pp. 77. Printed in Great Britain

BOOK REVIEW

J. R. O'Connor and J. Smiltens (Editors): **Silicon Carbide—A high temperature semiconductor.** (Proceedings of the Conference held at Boston, April 2–3, 1959). Pergamon Press, Oxford, 1960. 520 pp., 90*s*.

It is characteristic of present-day trends that a conference on the properties of one single substance should have attracted the attendance of more than 500 scientists and technologists, representing ten nations (though in overwhelming proportion from the U.S.). As indicated by the sub-title of the present volume, which contains the proceedings of the Conference, silicon carbide owes its importance to its potential usefulness as a high-temperature semiconductor (and one capable of resisting exposure to high nuclear radiation intensities for long periods of time). Such a semiconductor is urgently required, mainly for military applications, and that probably explains why one gets the impression, on reading these proceedings, of a "crash program", undertaken under direction for specific purposes, while on the other hand the devices which have caused so much research work to be undertaken are necessarily only vaguely hinted at. There is in fact only one paper describing a working device.

However, the properties of the substance, intermediate between silicon and diamond, are alone interesting enough. From a fundamental point of view they bring out well the essential differences between element and compound semiconductors, and the effect of a modification in crystal structure. For the technologist, it is not just a simple substitution of silicon carbide for Si or Ge in conventional diodes, transistors, etc., but the introduction of new device concepts, for the moment chiefly of military interest, but which may well break new ground in solid-state electronics.

The contents of the volume reflect faithfully the original conference programme. Section I deals with the thermodynamics of the silicon–carbon binary system (one paper considering also the ternary system C–Si–Ge), with discussions of the peculiarities of the phase diagram. Section II describes methods for growing SiC single crystals, of both the cubic and the hexagonal modifications, the method most fully considered being growth by gaseous cracking: four vapour systems have already been successfully applied. Section III, dealing with SiC as a solid, contains papers on its crystal structure, surface morphology, chemical imperfections and oxidation properties. Section IV is the one that will appeal most to research workers on semiconductors. After a brave attempt at unravelling the electronic band structures of the two modifications, both by fundamental and semi-empirical methods, the electrical properties are the subject of three papers, much interest being shown in the behaviour of SiC contacts at high temperatures. Then follow the optical properties which, as with most semiconductors, lead to the most unambiguous conclusions. A lot of optical research seems still to be necessary to determine the nature and effects of impurities. Finally, Section V is called "Silicon carbide devices", but, with the already quoted single exception, the title is somewhat misleading, since the section deals in fact with the processing of SiC and with preliminary measurements in view of applications.

At the end of the volume one still wonders: will silicon carbide fulfil its promise? One can only see that a great amount of research is still necessary before one can be in a position to make a well-founded assessment. The conference, and its proceedings, find their main usefulness in pointing out the most promising lines of research to achieve that goal. Apart from this, the book will be of great interest to all solid-state laboratories as a complete and authoritative review and discussion of the properties of a notable semiconductor. The Editors must be congratulated for their excellent and quick work, and the Publishers for the high standard of printing and photograph reproduction.

L. Pincherle

Solid-State Electronics Pergamon Press 1961. Vol. 2, p. 78. Printed in Great Britain

ERRATUM

L. A. D'ASARO: Diffusion and oxide masking in silicon by the box method. Vol. 1, No. 1, pp. 3–12. The first column in the last line of Table 1 (p. 9) should read "0·1 % P: 0·23% P_2O_5" and not "1% P: 23 % P_2O_5". The author apologizes for any inconvenience this may have caused readers.

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 79–84. Printed in Great Britain

THE HIGH-INJECTION-LEVEL OPERATION OF DRIFT TRANSISTORS

J. LINDMAYER and C. WRIGLEY

Sprague Electric Company, North Adams, Massachusetts

(*Received* 20 *October* 1960; *in revised form* 3 *November* 1960)

Abstract—Analysis of the minority-charge distribution in the base region of the drift transistor by a one-dimensional theoretical model results in an Abel-type differential equation which has no analytical solution. It is possible to compute the charge distribution, and, by normalization, it can be done with only two parameters for an exponential impurity distribution. The stored charge and the transit time are given as functions of the injection level. The overall time constant of the drift transistor shows a minimum at a particular current density; a good analytical approximation is given for this current density.

Résumé—L'analyse de la distribution des charges minoritaires dans la région de la base d'un transistor "drift" par un modèle théorique à une dimension résulte en une équation différentielle du type d'Abel n'ayant pas de solution analytique. Il est possible de calculer la distribution de charge, qui peut-être faite par normalisation avec deux paramètres, en considérant le cas d'une distribution exponentielle d'impuretés. La charge emmagasinée et la durée de transit sont présentées en fonction du niveau d'injection. La constante de temps totale du transistor "drift" démontre une valeur minimum à une certaine densité de courant; une bonne approximation analytique est donnée pour cette densité de courant.

Zusammenfassung—Die Untersuchung der Verteilung der Minoritätsträger in der Basiszone des Diffusionstransistors mit Hilfe eines eindimensionalen theoretischen Modells führt zu einer Abelschen Differentialgleichung, die keine analytische Lösung hat. Es besteht die Möglichkeit, die Ladungsverteilung zu berechnen, und zwar ist dies für eine exponentielle Störstellenverteilung durch Normalisierung mit nur zwei Parametern möglich. Die gestaute Ladung und die Laufzeit erscheinen als Funktionen des Injektionsniveaus. Die Zeitkonstante des Diffusionstransistors hat bei einer bestimmten Stromdichte ein Minimum; für diese Stromdichte wird eine gute analytische Annäherung gegeben.

SYMBOLS AND UNITS

J hole current density (A/cm^2)
q electronic charge (C)
μ mobility of holes (cm^2/V-sec)
D diffusion constant of holes (cm^2/sec)
E_b electric field in the base region (V/cm)
p hole concentration (cm^{-3})
n electron concentration (cm^{-3})
N impurity concentration (cm^{-3})
η dimensionless base width normalized to x_0
W base width (cm)
x_0 $1/e$ length of exponential (cm)
ξ dimensionless distance
Υ dimensionless hole concentration
δ dimensionless current density
τ_B base transit time (sec)
δ_{cr} critical dimensionless current density
V voltage in base region (V)
V_b voltage across base region (V)
p_E value of p at $x = 0$ (cm^{-3})
I_E emitter current (A)
ζ dimensionless factor
τ_E emitter-circuit time constant (sec)
τ_C collector time constant (sec)

THE properties of the homogeneously doped transistor's base region have already been calculated by SHOCKLEY[1] for small injection levels. RITTNER[2] has extended the theory for high in-

jection levels, and has shown that the electric field in the base region cannot be neglected.

In order to calculate the minority-carrier distribution, let us consider a narrow-base drift transistor, with an impurity distribution $N(x)$. Surface recombination, volume recombination and the equilibrium concentration of minority carriers are neglected, since their contributions are insignificant for the properties calculated. For convenience p–n–p transistors are considered.

The hole current may be written as

$$\frac{J}{q} = \mu E_b(x)p(x) - Dp'(x) \tag{1}$$

where the prime refers to a differential with respect to x.

Assuming a constant mobility, we can use the Einstein relation to write (1) as

$$\frac{J}{qD} = \frac{q}{kT}E_b p - p' \tag{2}$$

where kT/q is the thermal voltage.

If we now neglect the electron current, we may obtain an expression for the field by writing the electron current as

$$0 = \frac{q}{kT}E_b n + n' \tag{3}$$

where n is the concentration of electrons. The field is then

$$E_b = -\frac{kT}{q}\frac{n'}{n} \tag{4}$$

The space-charge neutrality in the base region requires that

$$n - N - p = 0 \tag{5}$$

Setting (5) into (4), the field obtained is

$$E_b = -\frac{kT}{q}\frac{(N+p)'}{N+p} \tag{6}$$

The impurity distribution, $N(x)$, is assumed to be exponential, since this closely describes the impurity distribution in drift transistors. For this distribution $N = N_E \exp(-x/x_0)$. Let us define η as

$$\eta = \frac{W}{x_0} = \ln\frac{N_E}{N_W} \tag{7}$$

where N_E is the impurity concentration at the emitter, and N_W is the impurity concentration at W. It can be seen that the electric field in a drift transistor is constant only for low injection levels, when $p \ll N$. For diffusion transistors, where $N_E = N_W$, the electric field is zero for $p \ll N$.

Combining (6) and (2), we obtain

$$p'N + 2pp' + pN' + p\frac{J}{qD} + N\frac{J}{qD} = 0 \tag{8}$$

This is an Abel-type differential equation of the second kind.* This equation does not have a known analytical solution. For the forthcoming treatment it is convenient to normalize it by defining the following quantities. Let us make a further assumption, i.e. W is current independent.

$$\xi = x/W \qquad \text{normalized distance} \tag{9}$$

$$\Upsilon(\xi) = p(x)\frac{qD}{JW} \qquad \text{normalized hole concentration} \tag{10}$$

$$\delta = \frac{JW}{N_E qD} \qquad \text{normalized hole-current density} \tag{11}$$

$$N = N_E e^{-\eta\xi} \qquad N \text{ as a function of } \eta \text{ and } \xi \tag{12}$$

The differential equation (8) may now be written in the form

$$\Upsilon' + 2\Upsilon\Upsilon'\delta e^{\eta\xi} - \eta\Upsilon + \Upsilon\delta e^{\eta\xi} + 1 = 0 \tag{13}$$

where the prime now refers to a differential with respect to ξ.

For small currents, when $\delta \ll 1/\Upsilon e^{\eta\xi}$, the solution of (13), subject to the boundary condition $\Upsilon = 0$ at $\xi = 1$, is

$$\Upsilon_{\delta\ \text{small}} = \frac{1 - \exp[-\eta(1-\xi)]}{\eta} \tag{14}$$

Of course, for extremely small currents the equilibrium concentration of holes cannot be neglected.

For large currents, when $\delta \gg 1/\Upsilon e^{\eta\xi}$, the solution becomes

$$\Upsilon_{\delta\ \text{large}} = \tfrac{1}{2}(1-\xi) + \frac{1}{2\delta}(e^{-\eta} - e^{-\eta\xi}) \tag{15}$$

The solid curves in Fig. 1 are for the low-current

* See, for example, MURPHY[3].

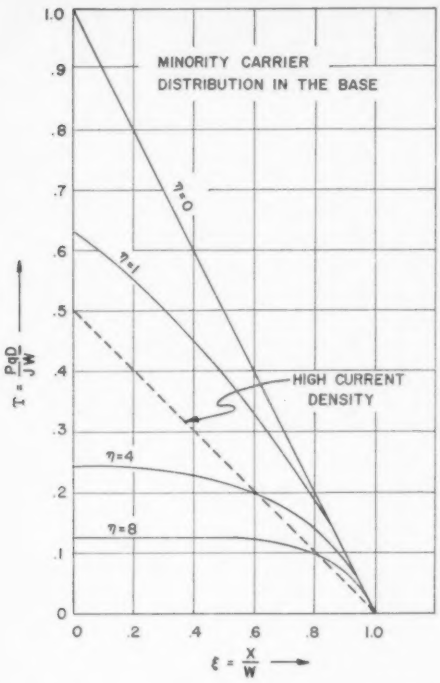


FIG. 1. The solid lines represent the hole distributions for low injection levels. As the current is increased indefinitely all distributions approach the dashed line.

solution, (14). For δ approaching infinity all solutions approach the dashed line in Fig. 1, which represents the first term in (15).

The full Υ-function was computed from $\xi = 1$ toward $\xi = 0$, with the initial condition $\Upsilon(\xi = 1) = 0$. In order to obtain good accuracy ξ was changed by steps of 0·01 until $\xi = 0{\cdot}8$, and thereafter by steps of 0·02. The values of Υ over the intervals were accumulated in order to obtain the definite integral for later use. The computed results for small δ's could hardly be distinguished from (14).

The transition of Υ from the low- to the high-current case is fairly simple for the diffusion transistor ($\eta = 0$), as Fig. 2 shows. This has already been given by RITTNER[2]. For drift transistors the transition is complicated, as is shown for $\eta = 8$ in Fig. 3. It is interesting to note the diffusion current flowing toward the emitter in a specific range of δ. Figs. 2 and 3 are partial results of the computation.

The transit time of injected carriers can be expressed from the total stored charge as has been shown by VARNERIN[4]. Therefore, we may write

$$\tau_B = \frac{q}{J}\int_0^W p\,dx = \frac{W^2}{D}\int_0^1 \Upsilon\,d\xi \qquad (16)$$

The definite integral of Υ · . (16) is (for ow currents), from (14),

$$\int_0^1 \Upsilon\,d\xi\Big]_{\delta\ \text{small}} = \frac{\eta - 1 + e^{-\eta}}{\eta^2} \qquad (17)$$

For $\eta = 0$ (diffusion transistor) (17) goes to 1/2, and therefore the transit time in (16) becomes the well-known $W^2/2D$. For large η (17) approaches

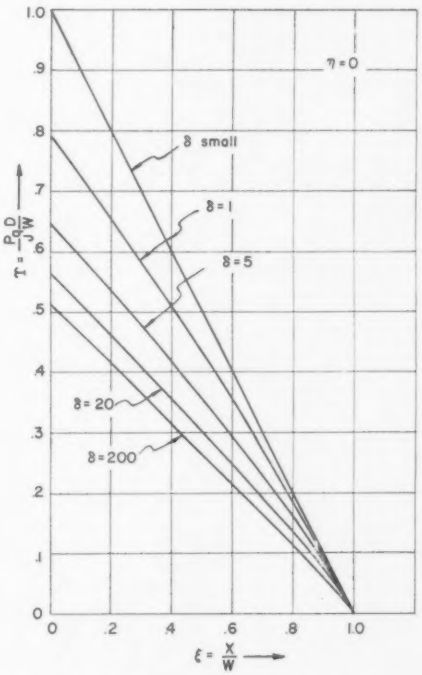


FIG. 2. The hole distribution of the diffusion transistor with increasing injection levels.

$1/\eta$, and the transit time is then $W^2/\eta D$. In the practical range $1/\eta$ is not a good approximation for (17), but $(\eta - 1)/\eta^2$ is quite good.

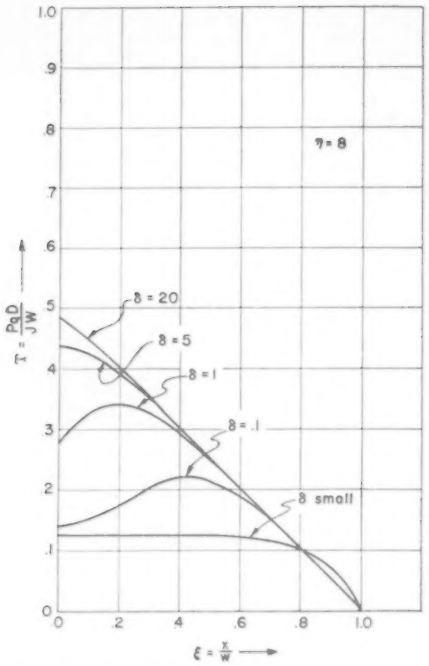


FIG. 3. The hole distribution for a drift transistor with $\eta = 8$ with increasing injection levels.

For high currents the definite integral of Υ from (15) is

$$\left[\int_0^1 \Upsilon \, d\xi\right]_{\delta \text{ large}} = \frac{1}{4} + \frac{1}{2\delta}\left(e^{-\eta} - \frac{1-e^{-\eta}}{\eta}\right) \tag{18}$$

For $\eta = 0$ this integral is 1/4, indicating that the high-current analytical solution results in the limiting case. The transit time is then $W^2/4D$, i.e. improved by a factor of 2. For large η the second term in (18) approaches $-1/2\delta\eta$, but in the limiting case as δ approaches infinity all transistors have the transit time $W^2/4D$.

In order to obtain the δ-dependence of the transit time, the definite integral needed for (16) was computed, and is plotted in Fig. 4. As can be seen, the improvement in a diffusion transistor's transit time with increasing δ is fairly slow.* However, the deterioration of τ_B for a good drift transistor ($\eta = 4$ to 8) occurs earlier. It may be of interest to note that a drift transistor with $\eta \simeq 2{\cdot}6$ will undergo no change in transit time over the whole range of δ.

The depletion layer at the emitter introduces a differential resistance and a capacitance, forming a time constant. For low currents the emitter differential resistance is kT/qI_E and the base region provides no contribution. However for high currents there is a contribution $(kT/qI_E)\zeta$ from the base region. ζ is a function which increases

* This was done for $\eta = 0$ by MEYER[5].

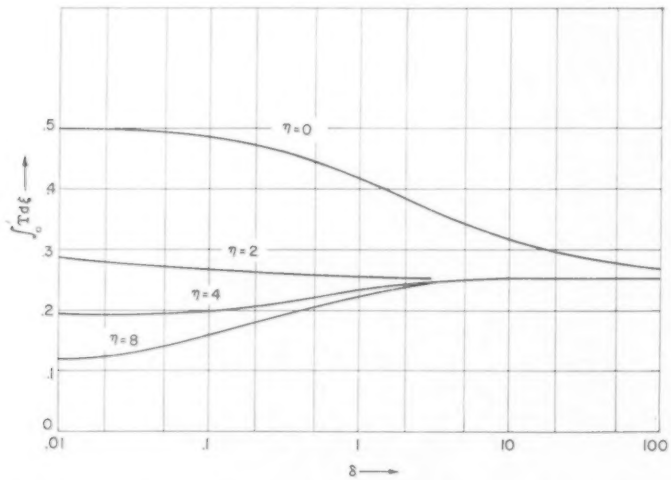


FIG. 4. Normalized stored charge as a function of η and the injection level.

from zero to unity as the current is increased. The emitter depletion-layer capacitance is very mildly dependent on the current, so we will assume that it is constant. The resulting time constant is

$$\tau_E = \frac{kT}{qI_E} C^* A_E (1+\zeta) \tag{19}$$

where A_E = emitter area and C^* = emitter depletion capacitance per unit area.

$$\frac{dV_b}{dI_E} = \frac{kT}{q} \frac{1}{N_E + p_E} \frac{dp_E}{dI_E}$$

$$= \frac{kT}{qI_E} \frac{\delta \Upsilon_E + \delta^2 (d\Upsilon/d\delta)_{\xi=0}}{\delta \Upsilon_E + 1} = \frac{kT}{qI_E} \zeta \tag{22}$$

where (10) and (11) were used for normalization. The actual ζ-function was computed and is plotted in Fig. 5. For $\eta \geq 2$ and $\delta < 1$, the following approximation fits the computed values

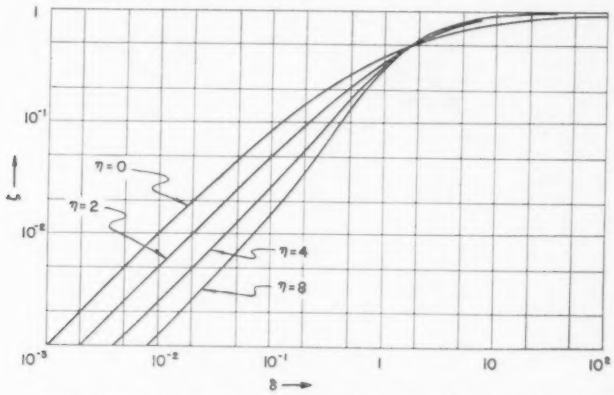


FIG. 5. The effective increment in kT/q as a function of η and the injection level.

This time constant improves with increasing δ, while in drift transistors τ_B deteriorates, resulting in a minimum over-all time constant at a particular current density. We will call this current density δ_{cr}. In order to calculate δ_{cr} we must know exactly how the emitter time constant changes with δ.

The voltage in the base region can be written as

$$V = -\int_W^x E_b \, dx = \frac{kT}{q} [\ln(N+p) - \ln N_W] \tag{20}$$

Since we have taken the collector space-charge boundary (W) as a reference, the voltage across the base region is

$$V_b = \frac{kT}{q} [\ln(N_E + p_E) - \ln N_W] \tag{21}$$

where $p_E = p(x=0)$. The base region's contribution to the differential input resistance is then

$$\zeta = \frac{\delta}{\eta} \tag{23}$$

For large δ and any η, using (15),

$$\zeta = \frac{\delta}{\delta + 1 + e^{-\eta}} \tag{24}$$

The over-all time constant may be written as

$$\tau = \tau_E + \tau_B + \tau_C \tag{25}$$

where τ_C, collector time constant, is δ independent.

The minimum transit time will occur when $d\tau/d\delta = 0$. Therefore from (25), using (11) and (16),

$$\frac{d}{d\delta} \int_0^1 \Upsilon \, d\xi = C_n \left[\frac{1}{\delta_{cr}^2} (1+\zeta) - \frac{1}{\delta_{cr}} \frac{d\zeta}{d\delta} \right] \tag{26}$$

where $d/d\delta$ is applied at $\delta = \delta_{cr}$, and

$$C_n = \frac{kT}{q} \frac{C^*}{N_E q W}$$

Negligible error results from use of (23) for the region $C_n \leqslant 2\times10^{-2}$, i.e. this requires that $\delta_{cr} < 1$, which is the range for practical drift transistors. The right-hand side of (26) then becomes C_n/δ_{cr}^2. The left-hand side was evaluated on a computer and is shown in Fig. 6. For $\delta < 1$, this function can be approximated within 5 per cent error by the analytic expression

$$\frac{d}{d\delta}\int_0^1 \Upsilon\, d\xi = \frac{\sinh\eta - 8}{2\sqrt{(2)}\eta^3(1+4\delta+0{\cdot}033\delta e^{0{\cdot}85\eta})} \tag{27}$$

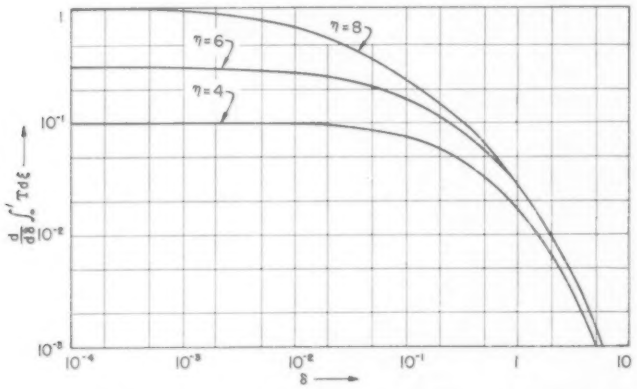


FIG. 6. The computed current derivative of the normalized stored charge as a function of the injection level.

Using (23) and (27) in (26), and denormalizing, we obtain

$$J_{cr} = \frac{N_E qD}{W} AB\left[1+\sqrt{\left(1+\frac{2}{B^2A}\right)}\right] \tag{28}$$

where

$$A = \frac{kTC^*}{q^2N_EW}\,\frac{\sqrt{(2)}\eta^3}{\sinh\eta - 8}$$

and $B = 4+0{\cdot}033\exp(0{\cdot}85\eta)$.
For $C_n \leqslant 10^{-3}$, (28) reduces to the simpler form

$$J_{cr} \simeq \frac{N_E qD}{W}\sqrt{(2A)} \tag{29}$$

The expressions above have been derived for a base width W which is independent of current. In the case of diffused collector junctions the effective base width may increase with current, resulting in a smaller J_{cr} and larger transit time at J_{cr}. However, for abrupt collector junctions the current dependence of W is small.

A brief comparison of the critical current density as defined in equations (28) and (29) with a critical current density as defined by KRÖMER[6]

$$J_{cr} = \frac{4qDN_E\eta e^{-\eta}}{W} \tag{30}$$

may be in order. KRÖMER uses the low-current solution for the minority-carrier distribution. He defines the critical current density when this distribution just reaches the impurity density at one point in the base region. This definition has no connection to specific electrical parameters, but was given as an approximation for the high current threshold. The critical current density derived here uses the applicable minority-carrier distribution and is related to the minimum transit time.

Acknowledgements—The writers are grateful to K. SCHOENI and L. WRIGHT of this company for mathematical advice and the computations.

REFERENCES

1. W. SHOCKLEY, *Bell. Syst. Tech. J.* **28,** 435 (1949).
2. E. S. RITTNER, *Phys. Rev.* **94,** 1161 (1954).
3. G. M. MURPHY, *Ordinary Differential Equations* p. 25. Van Nostrand, New York (1960).
4. L. J. VARNERIN, *Proc. I.R.E.* **47,** 523 (1959).
5. N. I. MEYER, *J. Electron. Control* **4,** 304 (1958).
6. H. KRÖMER, The drift transistor, *Transistors I* p. 215. RCA Laboratories (1956).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 85–95. Printed in Great Britain

DIFFUSION OF ARSENIC IN GERMANIUM FROM THE VAPOUR PHASE

W. ALBERS

Philips Research Laboratories, N.V. Philips' Gloeilampenfabrieken, Eindhoven, Netherlands

(*Received* 29 *July* 1960; *in revised form* 21 *November* 1960)

Abstract—The diffusion of arsenic in germanium has been measured. The diffusion constants and the concentrations of As at the surface in equilibrium with gaseous arsenic of a certain pressure have been determined at different temperatures. From the temperature dependence, the activation energy of diffusion and the pre-exponential constant have been calculated.

It is found that, if the diffusion is carried out via the vapour phase and if the diffusion vessel contains zones of different temperatures, the surface concentration of the arsenic depends on the geometry of the vessel, when working in the absence of an inert gas. This effect is caused by the fact that the degree of dissociation of the gaseous As molecules above the germanium sample depends on the geometry of the vessel.

A method is described how to get rid of thermal conversion.

Résumé—La diffusion de l'arsénic dans le germanium a été mesurée. Les constantes de diffusion et les concentrations d'As à la surface en équilibre avec l'arsénic gazeux à une certaine pression ont été déterminées à différentes températures. De cette dépendance de la température, l'énergie activant la diffusion et la constante pré-exponentielle ont été calculées.

On a trouvé que si la diffusion est obtenue en partant de la phase vapeur et si le vase de diffusion contient des zones de différentes températures, la concentration de surface de l'arsénic dépend de la géométrie du vase, quand l'opération se déroule en l'absence d'un gaz inerte. Cet effet est causé par le fait que le degré de dissociation des molécules d'As gazeuses au-dessus de l'échantillon de germanium dépend de la géométrie du vase.

On décrit une méthode pour éliminer la conversion thermale.

Zusammenfassung—Die Diffusion von Arsen in Germanium ist bereits untersucht worden. Die Diffusionskonstanten und die Konzentrationen von As auf der Oberfläche bei Gleichgewicht mit gasförmigem Arsen eines bestimmten Druckes sind bei verschiedenen Temperaturen bestimmt worden. Aus der Temperaturabhängigkeit wurden die Aktivierungsenergie für die Diffusion und die Diffusionskonstante vor dem exponentiellen Verlauf berechnet.

Wenn die Diffusion von der Dampfphase aus stattfindet, und das Diffusionsgefäss Zonen verschiedener Temperatur enthält, so hängt die Oberflächenkonzentration des Arsens von der Gestalt des Gefässes ab, wenn kein Edelgas vorhanden ist. Dies beruht auf der Tatsache, dass der Dissoziationsgrad der gasförmigen As-Moleküle über der Germaniumprobe von der Gestalt des Gefässes abhängig ist.

Es folgt die Beschreibung eines Verfahrens zur Verhinderung der thermischen Konversion.

1. INTRODUCTION

DIFFUSION of arsenic in germanium has been studied by FULLER[1], DUNLAP[2] and BÖSENBERG[3] by using the grinding method and also by following the displacement of p–n junctions. These were located by electrical probing. FULLER heated two low-resistivity germanium samples of different acceptor concentrations in contact with gaseous arsenic for one or more days. It was not reported whether the samples were immersed in gas or in vacuum, nor was the geometry of the diffusion vessel described. He assumed the validity of Fick's second law for constant diffusion coefficient.

DUNLAP heated the samples in vacuum and got fair agreement with theory for the dependence on the concentration of the diffusant on x (= distance

from the surface) and t (= time of diffusion). BÖSENBERG started with germanium single crystals in which steep impurity profiles were introduced by doping the melt when pulling the crystals. The deformation of the original profile by diffusion was derived from impedance measurements. There is no complete agreement between their results, as is shown in Table 1. The diffusion constant obtained by MCAFEE *et al.*[4] from impedance measurements is several orders of magnitude smaller.

The authors mentioned above[1–3] have measured the diffusion of As in Ge at rather high concentrations, the lowest lying at about 2×10^{15}–5×10^{15} As atoms cm^{-3}. To obtain a better accuracy it seemed worthwhile to measure the diffusion at lower concentrations. The displacement of p–n junctions in principle is the best method to study diffusion at low concentrations. However the well-known effect of thermal conversion plays a disturbing role in these low-concentration measurements. Recently LEVI[5] has found a means to avoid the thermal conversion under certain conditions. Starting from his observations it was possible to avoid thermal conversion in this kind of experiment too, and in this way to study diffusion at concentrations down to 10^{14} As atoms cm^{-3} or even lower. This will be dealt with in the Sections 2 and 3.

In Sections 4 and 5 the main body of the experimental material will be discussed. It will be shown that the diffusion of As at all concentrations and temperatures follows Fick's law. In all these experiments the pressure of the As is regulated by means of a cold zone in the diffusion vessel. It appeared, however, that, all other circumstances being similar, the concentration of the As at the surface of the germanium differed whether an inert gas was present or not. A number of possible explanations for this curious phenomenon are discussed, one of them pointing to the fact that, although in the experiments described the pressure of the arsenic may be determined, its temperature may differ for the two cases. The temperature is higher (and determined by the temperature of the germanium) in the case of the presence of an inert gas.

This hypothesis is confirmed by experiments described in Section 6. It will be shown that by a suitable form of the diffusion vessel, also in the case of the absence of an inert gas, the temperature of the gaseous As may be determined by the temperature of the germanium.

2. MATERIALS AND EXPERIMENTAL METHODS

The germanium crystals were p-doped with indium or gallium and grown or zone-levelled in the ⟨111⟩ direction. Slices (dimensions $1\times1\times0{\cdot}1$ cm^3) were sawn from these crystals. These were

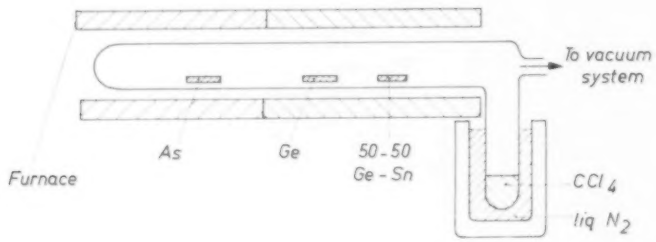


FIG. 1. Schematic drawing of the vacuum diffusion apparatus.

lapped on powder of silicon carbide and etched with a mixture containing 5 ml of concentrated nitric acid, 7 ml of 49 per cent hydrofluoric acid and 0·5 ml of ethanol, about 100 μ being taken off from each surface in order to ensure removal of any damaged layer. After etching, the etching liquid was diluted quickly by a large amount of deionized water. In this way a flat shining surface was obtained which was only slightly contaminated with copper.* The arsenic was highly purified by sublimation and was finely powdered before use.

The diffusion was carried out in a quartz tube

* Unpublished results of etching experiments on CP-4 etches contaminated with radioactive copper by Dr. B. OKKERSE and Mr. J. HORNSTRA.

having a diameter of about 2·5 cm (see Fig. 1). The tube was thoroughly cleaned with a mixture of nitric acid and hydrochloric acid, rinsed with deionized water and finally heated to temperatures close to the melting point. The quartz tube was situated in two electrically heated zones each 15 cm long, one for the germanium samples, the other for the arsenic.

In all experiments the arsenic was heated at 200°C, corresponding to a vapour pressure of about 10^{-4} mm Hg. The germanium was heated in successive experiments at different temperatures between 700 and 900°C. Both the arsenic powder and the horizontally placed slices of germanium were put in carbon boats which had been purified by heating them several hours at 1000°C, first in chlorine gas and then in vacuum. Quartz boats were also used. The diffusion time was mostly 2·25 hr. The temperature was measured with chromel–alumel thermocouples placed in a narrow quartz tube at the centre of the boats. The temperature was kept constant to within 5°C.

The residual air pressures in the vacuum experiments, maintained by an oil diffusion pump, were between 10^{-6} and 10^{-5} cm of Hg. In the gas experiments the flow velocity of the carrier gas was about 1 l./min, at a pressure of 1 atm. The concentration of the arsenic in the diffused layer was found by means of the *p–n*-junction displacement method.

The *p*-dope concentrations (specific resistance ranging from about 0·02 to 20 Ω-cm) were determined by four-probe conductance measurements, using tables correlating resistivity and carrier bulk concentration.[6]

The distance of the *p–n* junction from the germanium surface was measured with an accuracy of half a micron by means of an electrolytic copper-plating method.[7]

2.1. *Thermal conversion*

When germanium is heated at about 800°C, thermal conversion, i.e. a shift to *p*-type, generally occurs. The entire penetration curve of arsenic in germanium can be determined only by the *p–n*-junction method if thermal conversion is absent. Therefore one must get rid of this effect.

THEUERER and SCAFF[8] proposed two mechanisms for the thermal conversion, one postulating a change in solubility of an acceptor impurity, the other postulating the formation of acceptor centres by thermally produced lattice defects.

The experiments of FULLER and STRUTHERS[9], SLICHTER and KOLB[10], and SEILER *et al.*[11] indicate that the conversion is caused by quickly diffusing impurities, such as copper, which are present at the germanium surface before heating. VAN DER MAESEN *et al.*[12] have shown that thermal conversion is probably not caused by lattice defects arising from the germanium surface during heat treatment. These authors have also reported that germanium, converted into low-resistivity *p*-type by electroplating with copper and subsequent heating at 800°C, nearly recovers its original resistivity if heated at 500°C for at least 24 hr. This may be ascribed to the fact that the saturation concentration of copper is much smaller at 500°C than at 800°C, the values being in the order of 10^{13} cm^{-3} and 10^{16} cm^{-3} respectively.

KIKUCHI and IZIMA[13] have shown that removal of Cu from the bulk of germanium is not limited by the saturation concentration of Cu at the annealing temperature involved. From resistivity measurements they found that copper, introduced in germanium, can almost completely be removed by annealing several times at 650°C, interrupted by etching of the sample with concentrated nitric acid, which removes the copper from the surface. At the same time the original minority-carrier lifetime was practically restored. These experiments can be described by considering the surface and the bulk as a two-phase system and by introducing a distribution coefficient between the bulk concentration and the surface concentration of the copper.

When the specimens were immersed in CP-4 etch instead of in nitric acid the Cu atoms were not removed from the bulk of the germanium. KIKUCHI and IZIMA attributed this effect to the difference in roughness of the surfaces in the two cases. The experiments of KIKUCHI and IZIMA concerning thermal conversion are confirmed by our experiments, with the exception, however, that we found hardly any restoration of the lifetime. The residual copper concentration in the bulk, amounting to about $0{\cdot}2 \times 10^{14}$ cm^{-3}, is too low to explain the low ultimate lifetime of about 30 μsec. This may therefore be ascribed to another fast-diffusing impurity such as Ni. (10^{14}cm^{-3} Cu and 10^{14} cm^{-3} Ni correspond to a minority-carrier

lifetime of 180 μsec and about 1 μsec respectively.)[14,15] Annealing at another temperature results in another distribution of copper between bulk and surface. Our experiments show that the equilibrium between the copper at the surface and the copper in the bulk of the germanium shifts to the bulk side at increasing temperature. It is found that, relative to the same concentration of copper at the surface, the bulk concentration of copper at 800°C is about a factor of a hundred higher than that at 650°C.

KIKUCHI and IZIMA attributed the difference in their annealing experiments between germanium etched with nitric acid (rough) and CP-4 (shining) to the difference in surface properties. It should be realized, however, that the CP-4 etch does not easily dissolve copper from the germanium surface.* With regard to this, LEVI[5] described a method to "prevent" thermal conversion in shining-surface germanium during high-temperature vacuum diffusion using a 50–50 Sn–Ge alloy and frozen CCl_4 at liquid-nitrogen temperature. He suggested that the copper is removed from the germanium surface by means of chlorine gas coming from decomposing CCl_4, while the tin alloy may act as a copper sink. According to LEVI, a 37 Ω-cm *n*-type CP-4 etched germanium sample is found unchanged after the following procedure: heating in vacuum for 1 hr at 800°C, cooling slowly to 650°C (cooling rate of 2°C/min) and finally annealing for 1 hr at 650°C. This was confirmed by our experiments. LEVI also reports a conservation of minority-carrier lifetime, whereas we found a drop to 10 μsec and even lower.

If one introduces a temperature-dependent distribution coefficient of copper, as mentioned above, it becomes evident that LEVI actually did not prevent thermal conversion. Thermal conversion did occur at 800°C but he restored the original resistivity by cooling down slowly. In this way he obtained the best restoration velocity by combining the higher diffusivity of copper at higher temperatures with the higher surface concentrations of copper at lower temperatures. Therefore slowly cooling down may be expected to be favourable but not essential to the "prevention" of thermal conversion. The essential part of the procedure is the annealing at 650°C. This view is supported by our experiments which show thermal conversion if slow cooling is combined with annealing at a temperature higher than 650°C. If, on the other hand, a combination of rapid cooling and subsequent annealing at 650°C for a few hours is applied, no thermal conversion is observed. Thus it is possible to restore the original bulk resistivity of thermally converted germanium by annealing for some hours at 650°C in vacuum in the presence of frozen CCl_4 and 50–50 Sn–Ge alloy at 650°C. This is demonstrated by the experiments presented in Fig. 2.

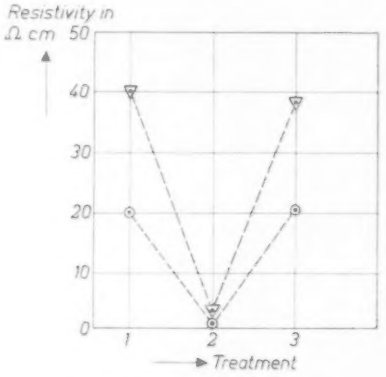


FIG. 2. Restoration of the original bulk resistivity. The vertical axis represents the specific resistance of two different germanium bars ($2 \times 0{\cdot}4 \times 0{\cdot}4$ cm^3) and the horizontal axis the number of treatment: (1) original resistivity of the germanium bars; (2) resistivity after immersion in a solution of copper nitrate in water and heating for 1 hr at 800°C; (3) resistivity of the bars after subsequent heating for 2–3 hr at 650°C in vacuum in the presence of CCl_4 at liquid-nitrogen temperature and a 50–50 Sn–Ge alloy at 650°C. An analogous behaviour was found if treatment number (2) was achieved by thermal conversion at 800°C without intentional contamination of the germanium surface with copper.

In our vacuum diffusion experiments the method of LEVI was followed. This method fails, however, in an inert-gas atmosphere. Then, to restore the original bulk resistivity, the germanium samples have to be annealed afterwards at 650°C in vacuum in the presence of frozen CCl_4 and a 50–50 Sn–Ge alloy for some hours. This can be done because the diffusivity of copper is much higher than the diffusivity of the conventional donors and acceptors.

* See footnote on p. 86.

3. MEASUREMENTS AND RESULTS

3.1. *Vacuum diffusion*

The germanium slices are clean in appearance but have a slightly matted surface after the diffusion period. The p–n barriers are exactly parallel to the flat surface. A typical example of the concentration of As during diffusion at various distances from the surface is shown in Fig. 3. The individual points in this figure were measured after a diffusion

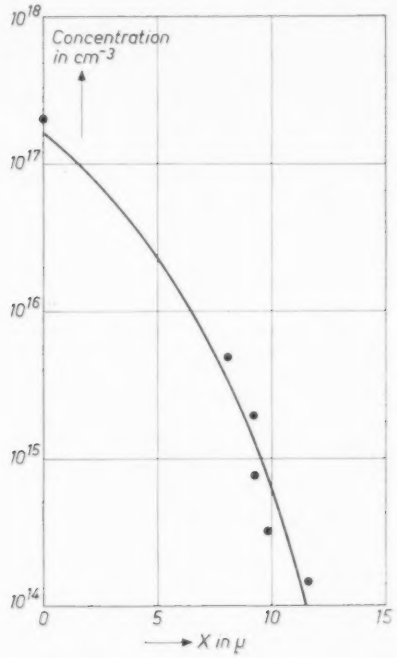


FIG. 3. Distribution of arsenic in germanium as a function of the distance x from the surface. The black points were found experimentally. The drawn curve corresponds to equation (1). The vertical axis represents the arsenic concentration. The arsenic temperature was 200°C and the germanium temperature 800°C. The diffusion time was 2·25 hr.

time of 2·25 hr, with the temperatures of the arsenic and germanium being 200 and 800°C respectively. These points are shown to be in reasonable agreement with the theoretical distribution (solid curve) according to the solution of Fick's second law for concentration-independent diffusion constant D and boundary conditions: $C(x, t) = 0$ (at $t = 0$ and $0 < x < \infty$) and $C(x, t) = C_e$ (at $x = 0$ and $0 < t < \infty$), reading

$$C(x, t) = C_e\left(1 - \operatorname{erf}\frac{x}{2\sqrt{(Dt)}}\right) \qquad (1)$$

where C_e represents the constant surface concentration of the diffusant and $C(x, t)$ the concentration in the solid at time t from the start of the diffusion at a distance x from the surface. The point at the vertical axis indicates that, at this original bulk concentration of the germanium slices, no p–n junction is observed, i.e. the surface concentration of the diffusant equals or is lower than the concentration indicated by this point. The diffusion constant D and the surface concentration C_e, obtained by extrapolation of the theoretical curve, are $0{\cdot}9 \times 10^{-11}$ cm^2 sec^{-1} and 2×10^{17} cm^{-3} respectively.

The temperature dependence of D is obtained from experiments in which the germanium was heated at 750°, 800°, 850° and 900°C. In Fig. 4,

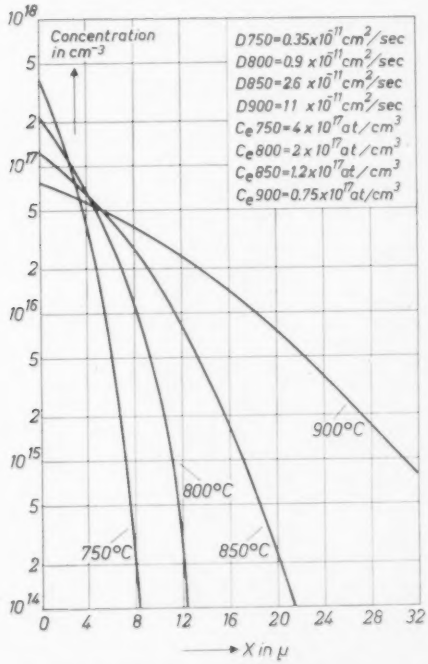


FIG. 4. The distribution of arsenic in germanium as a function of the distance x from the surface and as a function of the temperature of the germanium. The arsenic temperature was 200°C and the diffusion time was 2·25 hr.

theoretical distribution curves according to equation (1) are given, which show the best agreement with the experimental points measured at the different temperatures. In this way values of D are obtained for different temperatures which in Fig. 5 are plotted logarithmically against $1/T$. The straight dashed line through these points yields an activation energy of diffusion $E_{diff} = 56$ kcal/mole, and a pre-exponential constant $D_0 = 3$ cm^2 sec^{-1}. A comparison of our results with those of other investigators[1–3] is given in Table 1.

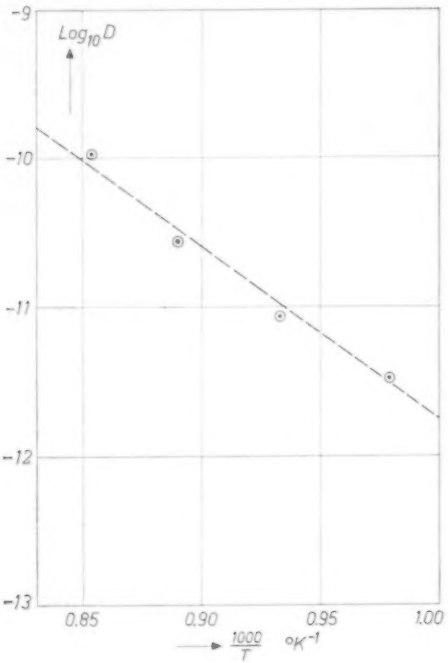


FIG. 5. Dependence on temperature of the diffusion constant of arsenic in germanium.

Table 1. Diffusion of arsenic in germanium in vacuum

D at 800°C (cm^{2}sec^{-1})	D_0 (cm^{2}sec^{-1})	E_{diff} (kcal mole^{-1})	Reference
4×10^{-11}	0·71	51	FULLER[1]
4×10^{-11}	10	57	DUNLAP[2]
1×10^{-11}	2·1	55	BÖSENBERG[3]
$1\times10^{-11}\pm15\%$	3	$56\pm7\%$	Our results

Fig. 6 shows diffusion patterns obtained after 57 hr (square points) and after 2·25 hr (circle points). In both cases arsenic and germanium were at 200 and 800°C respectively. Both distributions follow the lines drawn in accordance with equation

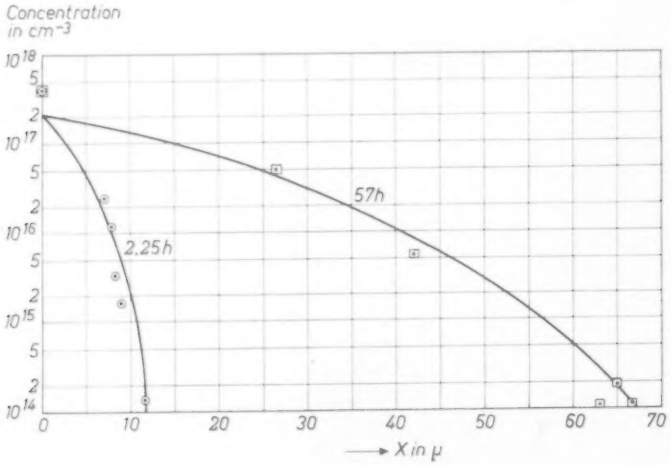


FIG. 6. The distribution of arsenic in germanium as a function of the diffusion time. The arsenic temperature was 200°C and the germanium temperature 800°C. The drawn curves correspond to equation (1).

(1). The same surface concentration is found in each case and also the diffusion constants agree satisfactorily. The values are $0{\cdot}9\times10^{-11}$ and $1{\cdot}0\times10^{-11}$ cm^2 sec^{-1} respectively.

3.2. *Diffusion in carrier gas*

Fig. 7 shows the distribution of arsenic in germanium after diffusion in the presence of argon, nitrogen or a mixture of 25 per cent hydrogen and

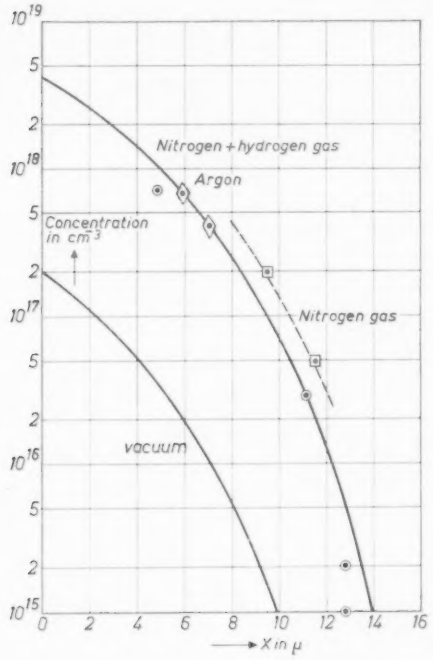


FIG. 7. The distribution of arsenic in germanium after diffusion in the presence of 1 atm argon (◇), nitrogen contaminated with oxygen (□) or a mixture of 25 per cent hydrogen and 75 per cent nitrogen gas (○). The arsenic temperature and the temperature of the germanium specimens were 200 and 800° respectively. The time of diffusion was 2·25 hr. The drawn curves correspond to equation (1).

75 per cent nitrogen gas (1 atm) compared with the one obtained in vacuum. Runs without arsenic showed that these gases contained only negligible amounts of donor impurities. Argon and the nitrogen–hydrogen mixture were purified by a copper pile to remove the oxygen. The nitrogen was not purified. Under the experimental conditions, the gases do not react chemically with the arsenic.[16] The diffusion time was 2·25 hr. The temperature of the arsenic source was 200°C and the germanium was at 800°C.

It is evident from this figure that the surface concentration is dependent on the ambient of the germanium samples. In the presence of an inert gas the diffusion pattern is shifted upwards as compared with the vacuum case. The surface concentrations in both cases differ by more than one order of magnitude, having values of 2×10^{17} cm^{-3} and 4×10^{18} cm^{-3} respectively. Again Fick's law is obeyed. The solid curves are the theoretical ones calculated from equation (1). The D values derived from both curves agree to within 20 per cent. The value is about 10^{-11} cm^2 sec^{-1}.

4. DISCUSSION

4.1. *Evaporation of the germanium surface*

Evaporation of germanium at the temperatures applied may be expected from known data to be negligibly small. Etching off of the germanium surface during diffusion treatment by inleaking oxygen from the air in the vacuum system might, however, be possible. Actually, after diffusion the germanium sample is observed to have a slightly matt surface.

By weighing the germanium specimens before and after blank diffusion treatment the rate of evaporation of germanium at the temperatures employed is found to be less than one tenth of a micron per hour. For this rate of evaporation, equation (1) is obeyed within a few per cent, i.e. within experimental error.

4.2. *Validity of Fick's law*

Fig. 6 illustrates the dependence of the arsenic penetration on time. Comparison of the two curves shows that over a wide concentration range the $\sqrt{(t)}$ relation is satisfied. At the same time it is shown that the short diffusion times adopted in almost all our experiments, i.e. 2·25 hr, give reliable results. This is partly due to the great accuracy of the measurements of the p–n-junction depths.[7] In Figs. 3, 6 and 7 the solid lines correspond to equation (1). The experimental points appear to scatter about these theoretical lines within quite narrow regions. Consequently the diffusion coefficient is roughly independent of the acceptor concentrations in the germanium specimens which

vary from 10^{14} to about 10^{18} cm^{-3}. This observation is supported by the experiments of DUNLAP[2], who found close agreement between the diffusion constants of antimony in germanium obtained by the p–n-junction method and the radioactive tracer method. This conclusion is noteworthy, the more so as BOLTAKS *et al.*[17] reported a variation of about a factor of 4 for the diffusion coefficient of antimony in germanium, with antimony dope concentrations ranging from 5×10^{14} to 5×10^{18} cm^{-3}.

5. CONCENTRATION AT THE SURFACE

The concentration at the surface of arsenic in germanium is an order of magnitude higher when diffused in inert gas ambient than when diffused in vacuum, as is demonstrated in Section 3.2. A discussion of several possible causes of this difference is given below.

5.1. *Oxidation of the diffusant*

The difference might be due to a different amount of oxygen leaking in during the vacuum experiments and during the inert-gas trials. The oxygen might chemically interfere with arsenic and might influence the composition of the gaseous arsenic to different extents. The results of the experiments described in Fig. 7, however, make the oxygen influence unlikely; approximately the same surface concentrations are obtained with a reducing and an oxygen-contaminated carrier gas as with a neutral gas. Evidently a specific influence of the carrier gas has to be ruled out.

5.2. *Surface barrier*

The solution of Fick's law leading to equation (1) starts from the pre-assumption that equilibrium surface concentration C_e is built up immediately. If, however, a surface barrier is present that has to be surmounted by the gaseous impurity atom in order to enter the germanium surface it may take a long time before the equilibrium surface concentration is reached. According to SMITS and MILLER[18], the boundary condition for the solution of Fick's second law in that case becomes

$$K[C_e - C(0,t)] = -D\left(\frac{dC(x,t)}{dx}\right)_{x=0}$$

where C_e represents the equilibrium concentration of the diffusant at the surface, $C(0,t)$ the surface concentration of the diffusant at time t [$C(0,t) \leqslant C_e$], D the concentration-independent diffusion constant and $C(x,t)$ the bulk concentration of the diffusant at distance x from the surface at time t. K is a rate constant. If we have the further boundary conditions: constant partial pressure of the gaseous arsenic and $C(x,t) = 0$ (at $t = 0$ and $0 < x < \infty$) the solution of Fick's second law becomes

$$\frac{C(x,t)}{C_e} = \left[1 - \operatorname{erf}\frac{x}{2\sqrt{(Dt)}} - \exp\left(\frac{K}{D}x + \frac{K^2}{D}t\right)\times \left\{1 - \operatorname{erf}\left(\frac{x}{2\sqrt{(Dt)}} + \frac{K}{D}\sqrt{(Dt)}\right)\right\}\right] \qquad (2)$$

(See Fig. 8).

According to equation (2) and Fig. 8, lengthening of the diffusion time by a factor of 25 must correspond to an increase in surface concentration

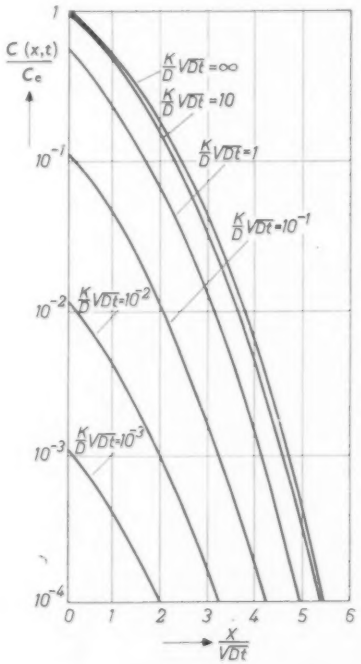


FIG. 8. Surface concentration gradually being built up. Theoretical distribution of arsenic in germanium as a function of the rate constant K, the diffusion constant D and the diffusion time t.

by a factor of 5 or a little less if non-equilibrium surface concentration is present in the vacuum diffusion experiment. The experiment of Fig. 6, however, shows that within experimental error the surface concentration does not change at all. Consequently, a difference in surface barrier between germanium in vacuum and germanium immersed in inert gas is not the reason for the difference.

Moreover, Fig. 6 shows that the surface barrier of germanium in vacuum, if present at all, is too small to prevent the establishment of equilibrium between arsenic dissolved in the germanium surface and gaseous arsenic within a relatively short diffusion time of 2·25 hr.

5.3. *Effect of pumping and cold spot*

In principle, the vapour pressure of the gaseous arsenic in the vacuum experiments will deviate from the maximum vapour pressure P_{max}. The transport of the gaseous diffusant will be provided by a free molecule flow[19] because the free-path length at the vapour pressures of about 10^{-4} mm Hg involved is considerably longer than the radius of the diffusion vessel.

It can be easily shown that

$$P_{stat}/P_{max} = 1 - \left(1 + \frac{O_e}{Qa^2} + \frac{O_e v_a}{O_c v_a + 4\beta}\right)^{-1} \quad (3)$$

where P_{stat} is the vapour pressure of the gaseous diffusant in the steady state right above the arsenic source, Q = Clausings factor,[20] O_e = evaporating arsenic surface, O_c = surface of the cold spot, v_a = average velocity of the gaseous diffusant molecules, β = pumping speed and a = radius of the tube in which the diffusion is carried out.

Considering our experimental conditions, it follows from equation (4) that the low surface concentration obtained in vacuum is not due to lowering of the vapour pressure of the arsenic by the influence of pumping or by the presence of a cold spot. The effect is small, even if O_c and β are large. (See Section 6 for experimental support of this conclusion.)

5.4. *The state of arsenic in the gaseous phase*

So far nothing has been said concerning the state of arsenic in the gaseous phase. Thermodynamic data on gaseous arsenic are unreliable and there is no conformity between experimental data. METZGER[20] reported the average number of atoms per molecule leaving the arsenic surface to be 2·6, whereas KANE and REYNOLDS[21] found no evidence whatsoever for species other than As_4 in the primary vapour of arsenic. The reported equilibria of dissociation do not agree either. It is evident, however, that dissociation will occur upon the transition of arsenic vapour from 200°C through a hot zone of 800°C. According to PREUNER and BROCKMÖLLER[22], mainly As_4 molecules and As atoms are present at 200°C and 800°C respectively at the pressure 10^{-4} mm of Hg applied.

Because arsenic in germanium is present as single atoms the equilibrium established refers to arsenic present in the vapour phase in atomic form. In the presence of an inert gas the arsenic molecules in the vicinity of the germanium sample collide practically exclusively with the hot inert gas molecules and therefore follow a Maxwell–Boltzmann velocity distribution corresponding to 800°C. In the vacuum case the arsenic molecules collide with both the cold and the hot walls of the vessel, and therefore in the steady state the velocity distribution will deviate from the one just mentioned. The velocity distribution in the vicinity of the germanium sample will shift to a region corresponding to a temperature somewhat lower than 800°C, which results in a diminished dissociation. Because of the high dissociation energy of arsenic, i.e. several electron volts, a considerable effect can be expected on the concentration of the arsenic atoms in the steady state. It may be markedly lower than if an inert gas is present. The difference in surface concentration observed after diffusion in vacuum and in carrier gas could be thought to be a demonstration of this effect.

6. GEOMETRY OF THE VESSELS IN THE VACUUM EXPERIMENTS

In order to investigate the effects mentioned in Sections 5.3 and 5.4 two further experiments have been carried out. In both, the diffusion of As was measured after 2·25 hr with the arsenic source at 200°C and the germanium sample at 800°C. Both experiments were carried out in vacuum. This time, however, the vacuum was not maintained by pumping; on the contrary, closed vessels were used that were evacuated and sealed before use. In one of the experiments the diffusion vessel consisted of a wide tube having a diameter of 2·5 cm. In the

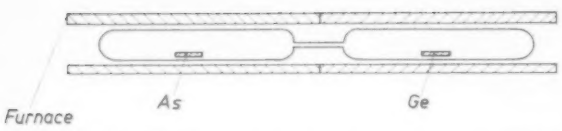


FIG. 9. Schematic drawing of the constricted tube.

other experiment the diffusion vessel was constructed in such a way that a wide compartment containing the arsenic source and a wide compartment containing the germanium sample were separated by a narrow tube having a diameter of 0·3 cm and a length of 5 cm (see Fig. 9). The result is shown in Fig. 10. The same diffusion constant is found (about 10^{-11} cm^2 sec^{-1}), but there is a marked difference in the surface concentration. The values are 5×10^{17} cm^{-3} for the wide tube and 4×10^{18} cm^{-3} for the constricted one. This difference is a proof that the difference between the surface concentrations in vacuum and in the presence of an inert gas is actually not mainly due to the effect of pumping or cold spot (Section 5.3).

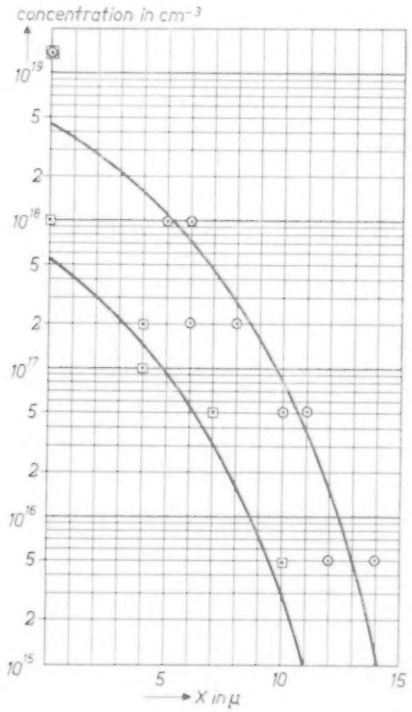


FIG. 10. The distribution of arsenic as a function of the distance x from the surface. The drawn curves correspond to equation (1). □—experiments carried out in a wide vacuum tube; ○—experiments carried out in a constricted tube.

The fact that the constricted tube shows the same surface concentration as was found in the previously described inert-gas experiments shows that the explanation given in Section 5·4 is the right one. By means of the constriction of the tube, the arsenic molecules in the vicinity of the germanium sample have lost almost every contact with the cold zone (the arsenic compartment). Therefore, also in this case, the temperature of the gaseous arsenic above the germanium is determined by that of the germanium sample and not partially by the temperature of the cold zone.

If one compares the surface concentration in the experiment with the wide, closed vacuum tube with the previous experiments with wide tubes and continuously pumping, the first one turns out to be somewhat higher. This also can be explained by the same arguments. Due to the impossibility of thoroughly outgassing the arsenic source the pressure in the closed system can easily rise to about 10^{-3}–10^{-2} mm Hg. This pressure is just sufficient to reduce the mean free path of the gaseous arsenic molecules to the dimensions of the diffusion vessel, that is to say that the influence of the cold zone will begin to be reduced.

7. CONCLUSION

Diffusion from a vapour phase often plays an important role. From the experiments just described it follows that:

(1) Fick's law is obeyed from concentrations of 5×10^{18} cm^{-3} down to 10^{14} cm^{-3}.

(2) If the diffusion vessel contains zones of different temperatures, the degree of dissociation of the gaseous arsenic molecules and hence the surface concentration depends on the geometry of the diffusion vessel, when working in the absence of an inert gas. Therefore, in order to obtain a definite concentration of a diffusant in a solid by means of a vacuum diffusion, a careful consideration of the geometry of the diffusion vessel is required. If not, an inaccuracy in the concentration of more than an order of magnitude may be expected.

(3) The surface concentration is independent of the geometry of the diffusion vessel when the diffusion is carried out in an inert-gas atmosphere.

Acknowledgements—The author wishes to thank Mr. J. GOORISSEN and Mr. P. C. VAN DER LINDE for the supply of the germanium crystals and purified arsenic. Mr. B. JANSEN for the measurements of the diffusion depths and Dr. F. VAN DER MAESEN and Dr. H. J. VINK for useful discussions.

REFERENCES

1. C. S. FULLER, *Phys. Rev.* **86,** 136 (1952).
2. W. C. DUNLAP, JR., *Phys. Rev.* **94,** 1531 (1954).
3. W. BÖSENBERG, *Z. Naturf.* **10,** 285 (1955).
4. K. B. MCAFEE, W. SHOCKLEY and M. SPARKS, *Phys. Rev.* **86,** 137 (1952).
5. C. A. LEVI, Communicated at the *AIEE-IRE Semiconductor Devices Research Conference, Boulder, Colorado, July,* 1957.
6. M. B. PRINCE, *Phys. Rev.* **92,** 681 (1953); R. N. HALL, *J. Phys. Chem. Solids* **3,** 72 (1957).
7. B. JANSEN, To be published.
8. H. C. THEUERER and J. H. SCAFF, *J. Metals, N.Y.,* **191,** 59 (1951).
9. C. S. FULLER and J. D. STRUTHERS, *Phys. Rev.* **87,** 526 (1952).
10. W. P. SLICHTER and E. D. KOLB, *Phys. Rev.* **87,** 527 (1952).
11. K. SEILER, D. GEIST, K. KELLER and K. BLANK, *Naturwissenschaften* **40,** 56 (1953).
12. F. VAN DER MAESEN, P. PENNING and A. VAN WIERINGEN, *Philips Res. Rep.* **8,** 241 (1953).
13. M. KIKUCHI and S. IZIMA, *J. Phys. Soc. Japan* **12,** 824 (1957).
14. J. A. BURTON, G. W. HULL, F. J. MORIN and J. C. SEVERIENS, *J. Phys. Chem.* **59,** 853 (1953).
15. G. K. WERTHEIM, *Phys. Rev.* **115,** 37 (1959).
16. W. T. COOKE, *Proc. Roy. Soc.* A**77,** 148 (1906); A. C. VOURNASOS, *Ber. dtsch. Chem. Ges.* **43,** 2264 (1910).
17. B. I. BOLTAKS, V. M. PROKHOROVA and L. I. NOVOZHILOVA, *Soviet Phys. Tech. Phys.* **3,** 92 (1958).
18. F. M. SMITS and R. C. MILLER, *Phys. Rev.* **104,** 1242 (1956); **107,** 65 (1957).
19. S. DUSHMAN, *Scientific Foundations of Vacuum Techniques* p. 90. John Wiley, New York (1949).
20. F. METZGER, *Helv. phys. acta* **16,** 323 (1943).
21. J. S. KANE and J. H. REYNOLDS, *J. Chem. Phys.* **25,** 342 (1956).
22. G. PREUNER and I. BROCKMÖLLER, *Z. phys. Chem.* **81,** 129 (1912).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 96–99. Printed in Great Britain

ELECTRICAL AND OPTICAL PROPERTIES OF $CdSnAs_2$

W. G. SPITZER, J. H. WERNICK and R. WOLFE

Bell Telephone Laboratories, Murray Hill, New Jersey

(*Received* 12 *September* 1960)

Abstract—The electrical and optical properties of *n*-type $CdSnAs_2$ have been investigated for samples having carrier concentrations between 1 and 3×10^{18} cm^{-3}. Resistivity and Hall coefficient measurements yield room-temperature electron Hall mobilities between 6200 and 5500 cm^2/V-sec in this carrier concentration range. Measurements between room temperature and liquid-helium temperature show degenerate behavior. Above room temperature the beginning of the intrinsic range was observed with an approximate energy gap of 0·3 eV. A thermal-conductivity measurement yielded a phonon thermal conductivity of 0·071 W/cm°C. For the same carrier concentrations, the thermoelectric power ranged from -100 to -70 μV/°C, indicating a thermoelectric figure of merit of $\sim 10^{-4}$/°C. Room-temperature infrared reflectivity and transmission measurements were made on the same samples used to obtain the electrical data. An electron effective mass between 0·04 m_0 and 0·06 m_0 and a lattice dielectric constant of $12{\cdot}1 \pm 0{\cdot}4$ were deduced from the reflectivity data. The shift in the absorption edge with increasing carrier concentration and the free-carrier absorption yielded electron effective masses in qualitative agreement with the values obtained from the reflectivity data.

Résumé—Les propriétés électriques et optiques de $CdSnAs_2$ du type *n* ont été examinées en employant des échantillons ayant des concentrations de porteurs allant de l à 3×10^{18} cm^{-3}. Les mesures de résistivité et du coefficient de Hall produisent, à température ambiante, des mobilités Hall d'électrons de 5500 à 6200 cm^2/V-sec dans cette gamme de concentration de porteurs. La différence entre les mesures à température ambiante et celles à température de helium liquide dénote un comportement dégénéré. Au-dessus de la température ambiante, le commencement de la gamme intrinsèque a été observé et un intervalle d'énergie d'environ 0,3eV noté. La valeur de la conductibilité thermique de phonon mesurée en employant des méthodes de conductibilité thermique est de 0,071 W/cm°C. Pour les mêmes concentrations de porteurs, la puissance thermoélectrique a été de -100 à -70 μV/°C indiquant un facteur de mérite thermoélectrique d'environ 10^{-4}/°C. Mesures de réflectivité infra-rouge et de transmission ont été faites à température ambiante sur les mêmes échantillons employés pour obtenir les données électriques.

Par ces données de réflectivité, on a déduit pour la masse effective d'électrons des valeurs allant de $0{,}04m_0$ à $0{,}06m_0$ et pour la constante diélectrique de reseau une valeur de $12{,}1 \pm 0{,}4$. Le déplacement de l'arête d'absorption en fonction de l'augmentation de la concentration des porteurs et l'absorption de porteurs libres ont produit des masses effectives d'électrons en accord qualitatif avec les valeurs obtenues par les données de réflectivité.

Zusammenfassung—Die elektrischen und optischen Eigenschaften von *n*-Typ $CdSnAs_2$ wurden an Proben mit Trägerkonzentrationen von 1 bis 3×10^{18} cm^{-3} untersucht. Messungen von spezifischen Widerstand und Hallkoeffizient ergeben bei Zimmertemperatur eine Elektronen Hall-Beweglichkeit von 6200 bis 5500 cm^2/V-sec innerhalb dieses Bereichs der Trägerkonzentration. Messungen zwischen Zimmertemperatur und der Temperatur von flüssigem Helium zeigen Entartung. Oberhalb von Zimmertemperatur begann der Eigenbereich mit einer Bandabstand von etwa 0,3 eV. Messung der Wärmeleitfähigkeit gab eine Phononenleitfähigkeit von 0,071 W/cm°C. Für dieselbe Trägerkonzentration lag die Thermokraft zwischen -100 bis -70 μV/°C, so dass die thermoelektrischen Leistungsziffer etwa 10^{-4}/°C beträgt. Dieselben Proben wurden bei Zimmertemperatur auf ihr Reflexionsvermögen und Transmissionsvermögen für Infrarot untersucht. Eine effektive Elektronenmasse zwischen 0,04 m_0 und 0,06 m_0 und eine Gitter dielektrischen Konstante von $12{,}1 \pm 0{,}4$ wurden aus den Reflexionseigenschaften abgeleitet. Unter Benutzung der Verschiebung des Absorptionsrandes mit zunehmender Trägerkonzentration und der Absorption der freien Träger erhielt man effektive Elektronenmassen, die mit den Ergebnissen der Reflexionsuntersuchungen qualitativ übereinstimmten.

1. INTRODUCTION

$CdSnAs_2$ is one of a number of ternary semiconductor compounds[1] containing Periodic Groups 2, 4 and 5 elements. It has the Chalcopyrite structure[1, 2], and the co-ordination is tetrahedral. It is an analogue of InAs with alternate atoms of In replaced by Cd and Sn. The tetragonal unit cell can be considered as two unit cells of InAs ($c/a = 2$). Some room temperature Hall mobility (μ_H) measurements have been reported for this material. FOLBERTH and PFISTER[1] gave $\mu_H = 3000$ cm^2/V-sec for an n-type sample where the carrier concentration was unspecified. A similar measurement recently reported by ROSENBERG and STRAUSS[3] gave $\mu_H = 5600$ cm^2/V-sec for $n = 2{\cdot}7 \times 10^{18}$ cm^{-3}. They also reported[4] that a sample with $n = 5 \times 10^{17}$ cm^{-3} had a $\mu_H = 12{,}000$ cm^2/V-sec. Preliminary optical-transmission data mentioned by these authors[4] indicated a minimum in the absorption coefficient near 5μ.

2. EXPERIMENTAL METHODS AND RESULTS

In the present study, polycrystalline ingots of $CdSnAs_2$ were prepared by melting stoichiometric amounts of the elements in graphite boats sealed in quartz tubes. The Cd and Sn were of 99·9999 per cent purity. The As (99·999 per cent pure) was heated in vacuum prior to use, in order to remove oxygen. The maximum preparation temperature was 750°C (melting point estimated to be ~600°C) and the melts were allowed to cool to room temperature at rates from 5 to 40°C/hr. The cooling was controlled to minimize cracking which is normally associated with cast ingots. The grain size in these ingots was of the order of 0·1 cm. Powder X-ray patterns did not indicate the presence of any second phase.

Electrical measurements were made by using the method described by VAN DER PAUW[5] on "Maltese cross" shaped samples. The significant center portion of each specimen was a circular region 0·16 cm in diameter and 0·05 cm thick. The resistivity ρ, Hall coefficient R and Seebeck coefficient S were measured on three specimens at room temperature and the results are given in Table 1. The conduction-electron concentration N is assumed to be given by $N = -1/Re$, and the Hall mobility μ_H is defined as $|R|/\rho$.

Measurements of ρ and R as functions of temperature between 295°K and 4·2°K were made on samples nos. 1 and 3. In each case the Hall coefficient remained constant over the entire temperature range. The resistivities decreased by ~10 per cent between 295°K and 77°K and then remained constant to 4·2°K. The low-temperature behavior is characteristic of a degenerate semiconductor in which impurity scattering predominates. On a sample with properties similar to those of no. 1, measurements of R and ρ were made at temperatures above 295°K. R remained constant and ρ increased slightly with temperature to 470°K. When the temperature was increased beyond 470°K both R and ρ decreased indicating the onset of the intrinsic range. While the activation energy appeared to be ~0·3 eV, an accurate value could not be determined because of thermal decomposition at these elevated temperatures. Irreversible decreases in both R and ρ were introduced when the specimen was raised to temperatures >720°K.

The thermal conductivity was measured at 290°K by an absolute (Lee's disk)[6] method on a sample where $\rho = 3{\cdot}3 \times 10^{-4}$ Ω-cm. The value obtained was 0·092 W/cm°C. Assuming a degenerate electron distribution, the electronic contribution was estimated to be ~0·021 W/cm°C yielding a phonon thermal conductivity K_{ph} of ~0·071 W/cm°C.

Room temperature infrared reflectivity measure-

Table 1. *Properties of* $CdSnAs_2$ *at* 295°K

Specimen	ρ (Ω-cm)	R (cm^3/C)	N (cm^{-3})	μ_H (cm^2/V-sec)	S (μV/°C)
1	$8{\cdot}9 \times 10^{-4}$	−5·5	$1{\cdot}1 \times 10^{18}$	6200	−100
2	$5{\cdot}0 \times 10^{-4}$	−2·9	$2{\cdot}1 \times 10^{18}$	5900	− 85
3	$3{\cdot}6 \times 10^{-4}$	−2·0	$3{\cdot}1 \times 10^{18}$	5500	− 70

ments of the central region of the three samples in Table 1 were made by using the method described by SPITZER and FAN[7]. The results between 5 and 30 μ wavelength are shown in Fig. 1. The electron effective mass m^* can be estimated[7]

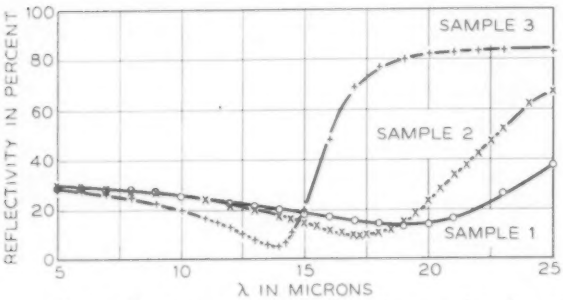


FIG. 1. Room-temperature reflectivity as a function of wavelength for the three samples of Table 1.

from the free-carrier electric susceptibility determined from the reflectivity measurements. The reliability of this type of measurement has been established for several other semiconductors.[7, 8] However, in order to determine the susceptibility, it is necessary to know ϵ_0, the dielectric constant in the absence of free carriers. Since samples of low-carrier concentration were not available, various values for ϵ_0 were assumed and the susceptibility calculated. Between 5 and 20 μ the value of ϵ_0 which yields the theoretical λ^2 dependence of the susceptibility is $12{\cdot}1 \pm 0{\cdot}4$. With this value for ϵ_0, the m^* values are between $0{\cdot}04\ m_0$ and $0{\cdot}06\ m_0$ for the three samples, where m_0 is the free-electron mass. The results are not sufficiently accurate to indicate whether m^* is a function of N. Also, since $CdSnAs_2$ is not a cubic crystal, it may be electrically and optically anisotropic, and the data obtained would be an average over different crystallographic directions.

Transmission measurements of samples nos. 1 and 2 were used with the reflectivity data of Fig. 1 to calculate the absorption coefficients, and the results are shown in Fig. 2. The shift in the intrinsic absorption edge due to filling of states at the bottom of the conduction band in samples of high-carrier concentration[9] is clearly visible near $3{\cdot}5$ μ. The presence of this shift is in qualitative agreement with the small effective mass found from the reflectivity data. Attempts to determine m^* and the position of the absorption edge in a non-degenerate sample from the observed shift cannot be made unless the value of m^* as a function of carrier concentration is known. For example, if $m^* \sim 0{\cdot}04\ m_0$ and independent of N, then sample no. 1 is $\sim 0{\cdot}2\ kT$ degenerate at room temperature, and the position of its edge is close

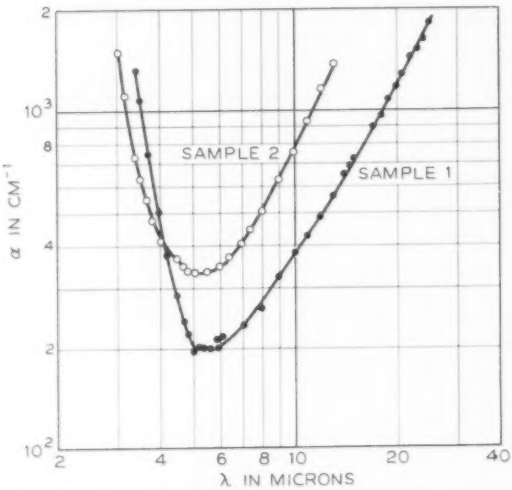


FIG. 2. Absorption coefficient as a function of wavelength for samples 1 and 2 of Table 1.

to that of a pure sample. If this is the case, then the threshold of the absorption edge at $\sim 5\ \mu$ or $\sim 0{\cdot}25$ eV is in agreement with the approximate value for the energy gap ($0{\cdot}3$ eV) observed in the high-temperature electrical measurements. However, if, as in InSb[7], m^* is considerably smaller at low carrier concentrations, then sample no. 1 could be several kT degenerate and the observed edge would be quite different in its spectral position from that of a pure specimen. Also in view of the possible optical anisotropy previously mentioned, any such calculation should certainly be regarded with suspicion. The free-carrier absorption for $\lambda > 5\ \mu$ shows close to a λ^2 dependence. However, elementary theory[7] predicts that the conductivity σ should be proportional to λ^2, and the conductivity is related to the absorption coefficient by $\alpha = 4\pi\sigma/nc$, where n is the refractive index. Since the reflectivity measurements indicate n is decreasing with increasing λ in the range used for the absorption measurements, σ will have a

wavelength dependence somewhat less than λ^2. The electron effective mass may be roughly estimated by using the free-carrier absorption and the Drude–Kronig[10] formula and is in reasonable agreement with the values deduced from the reflectivity.

3. CONCLUSIONS

The values of μ_H and m^* found for $CdSnAs_2$ are similar to those observed for InAs[7, 11] at the same electron concentrations. The lattice thermal conductivity, however, is about four times smaller than that of InAs.[12] The thermoelectric "figure of merit", $z = S^2/K\rho$, of $CdSnAs_2$ is low at room temperature ($z \sim 10^{-4}/°C$). However, if alloys of $CdSnAs_2$ with InAs can be made, they might be expected to have electrical properties similar to the end members[13], but their thermal conductivities should be greatly reduced. The thermoelectric figure of merit would therefore be considerably higher than either of the compounds (cf. InAs–GaAs alloys[12]).

Acknowledgement—The experimental assistance of D. Dorsi, S. E. Haszko and L. E. Howarth is gratefully acknowledged.

REFERENCES

1. O. G. Folberth and H. Pfister, *Semiconductors and Phosphore*. Interscience, New York (1958).
2. H. Pfister, *Acta Cryst.* **11,** 221 (1958).
3. A. J. Rosenberg and A. J. Strauss, *Bull. Amer. Phys. Soc.* **5,** 83 (1960).
4. This measurement was reported verbally at the New York meeting of the *Amer. Phys. Soc.*, January 1960.
5. L. J. van der Pauw, *Phillips Res. Rep.* **13,** 1 (1958).
6. H. J. Goldsmid, *Applications of Thermoelectricity* p. 78. Methuen, London (1960).
7. W. G. Spitzer and H. Y. Fan, *Phys. Rev.* **106,** 882 (1957).
8. W. G. Spitzer and J. M. Whelan, *Phys. Rev.* **114,** 59 (1959).
9. E. Burstein, *Phys. Rev.* **93,** 632 (1954).
10. H. Y. Fan and M. Becker, *Semiconducting Materials*. Butterworths, London (1951).
11. E. S. Greiner. Private communication.
12. M. S. Abrahams, R. Braunstein and F. D. Rosi, *J. Phys. Chem. Solids* **10,** 204 (1959).
13. A. F. Ioffe and L. S. Stil'bans, *Rep. Progr. Phys.* **22,** 167 (1959).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 100–105. Printed in Great Britain

COMPOUND CRYOSARS FOR LOW-TEMPERATURE COMPUTER MEMORIES*

ROBERT H. REDIKER
and
ALAN L. McWHORTER†

Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Massachusetts

(*Received* 12 *September* 1960)

Abstract—The compound cryosar, which operates at liquid-helium temperature, consists of two regions of compensated germanium, one more heavily doped than the other, in series between two ohmic contacts. When the more lightly doped region is broken down by impact ionization of the impurities, the impurity conductance of the more heavily doped region limits the current. This current limiting, in addition to the negative resistance of the breakdown, permits one to use a matrix array of compound cryosars in a coincident-voltage memory. If desired, the impact ionization breakdown of the more heavily doped region may be used for a non-destructive readout. High-density arrays of compound cryosars can be operated in a high-speed (10–100 Mc/s) computer memory, with a power dissipation for 200,000 elements of less than 0·25 W.

Résumé—Le cryosar maclé, qui fonctionne à la température du helium liquide, consiste en deux régions de germanium compensé, l'une plus fortement dopée que l'autre, en série entre deux touches ohmiques. Quand la région plus légèrement dopée est rompue par l'ionisation de choc des impuretés, la conductance d'impuretés de la région plus fortement dopée limite le courant. Cette limitation de courant et la résistance négative issue de la rupture permettent l'emploi d'un assemblage matriciel de cryosars maclés dans une mémoire à voltage coincident. Si on le désire, on peut utiliser la rupture des régions fortement dopées par choc d'ionisation pour une lecture non-destructuve de la mémoire.

Des assemblages à forte densité de cryosars maclés peuvent être employés dans la mémoire des calculateurs électroniques de grande vitesse (10–100 Mc/s) en ne dissipant que 0,25 W de puissance pour 200 000 éléments.

Zusammenfassung—Der zusammengesetzte Cryosar, der bei der Temperatur von flüssigem Helium arbeitet, besteht aus zwei Zonen von kompensiertem Germanium, von denen die eine stärker dotiert ist als die andere. Sie sind zwischen zwei Ohmschen Kontakten hintereinandergeschaltet. Wenn die schwächer dotierte Zone durch Stossionisation der Störstellen zusammenbricht, begrenzt die Störstellenleitfähigkeit der stärker dotierten Zone den Strom. Diese Strombegrenzung zusammen mit dem durch den Zusammenbruch entstandenen negativen Widerstand gestattet die Verwendung einer Matrixanordnung von zusammengesetzten Cryosaren in einem Speicher mit Spannungskoinzidenz. Es ist auch möglich, den Zusammenbruch der stärker dotierten Zone bei Stossionisation für eine zerstörungsfreie Lesung zu benutzen. Dichte Anordnungen von Cryosaren lassen sich als Speicher eines Rechners hoher Frequenz (10–100 MHz/sec) verwenden. Der Energieverlust für 200 000 Elemente beträgt weniger als 0,25 W.

---

* The work reported in this paper was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology with the joint support of the U.S. Army, Navy and Air Force. A preliminary report of this work was presented at the IRE-AIEE Solid-State Device Research Conference, Cornell University, June, 1959.

† Also at the Electrical Engineering Department, Massachusetts Institute of Technology, Cambridge, Massachusetts.

INTRODUCTION

COMPUTER components based on impact ionization[1] of impurities in germanium at liquid-helium temperatures have previously been described.[2] These components, named cryosars, exhibit a high resistance until a critical electric field is reached, after which the current increases by many orders of magnitude. If compensated germanium is used, a negative-resistance region occurs between the high- and low-impedance states, as shown in Fig. 1, permitting bistable operation. The voltage–current characteristics of Fig. 1 are for two samples of indium-doped germanium, compensated with antimony, in the form of wafers with ohmic contacts. The characteristics are the same in the third quadrant as in the first.

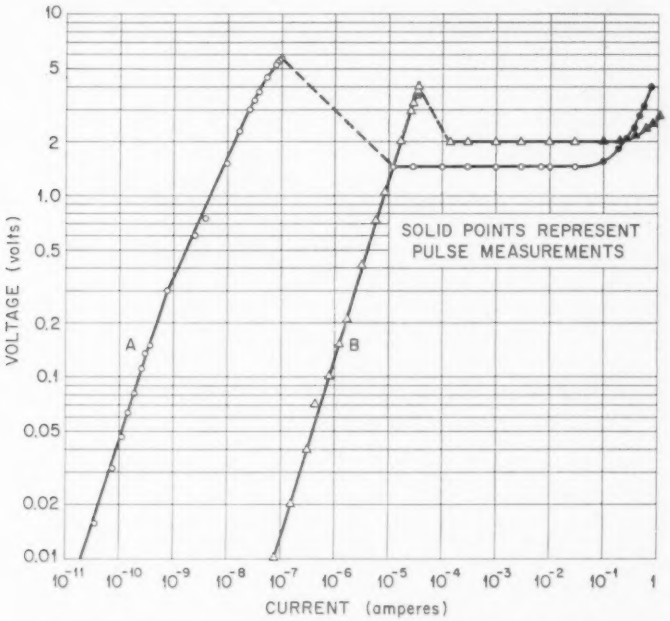


FIG. 1. Voltage–current characteristics at 4·2°K of indium-doped germanium strongly compensated with antimony. Sample *B* is more heavily doped than *A*. The data are for a sample thickness of 0·043 cm and a contact diameter of 0·1 cm.

A mechanism for this negative resistance has recently been proposed[3] involving the inelastic scattering of free carriers by pairs of nearby majority impurities, which will form a configuration analogous to a hydrogen molecule-ion if singly ionized by compensation. This extra scattering process disappears after breakdown since the "molecules" become fully ionized, allowing the breakdown to be sustained at a lower field than that required for its initiation. It has been found experimentally that the breakdown occurs in a thin filament.

MATRICES OF CRYOSARS

The localization of the breakdown permits many independent cryosars to be placed on one wafer of germanium. The yield problems usually encountered in semiconductor junction devices are not present, since the cryosar properties are relatively surface independent and all contacts are ohmic.[2] Reproducibility has seemed to be almost entirely a matter of controlling the impurity densities, which is not difficult on account of the techniques developed for transistor production. The matrix configuration shown in the sketch of Fig. 2 is particularly suitable for a computer

memory. The active regions of the cryosars are at the projected intersections between the ohmic line contacts to the top face and the ohmic line contacts to the bottom face. Fig. 3 shows a 32×32 matrix containing 1024 bistable cryosars in an

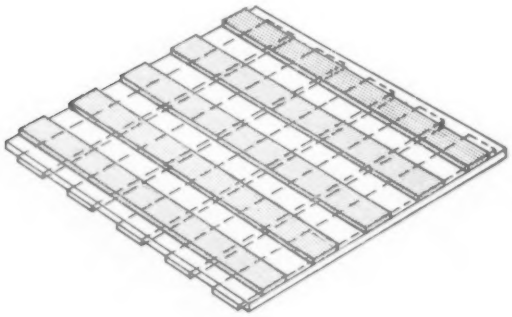


FIG. 2. Sketch of a matrix of twenty-five cryosars.

area of $\frac{1}{4}$ in^2. In the figure no connections are shown to the ohmic line contacts, which are evaporated aluminum, but in other matrices the connections have been made by thermocompression bonding of 2–3 mil gold wire. One would like to use such matrices in a coincident-voltage memory in which the state of a particular cryosar is changed when, and only when, pulses of appropriate magnitude and polarity are applied in coincidence to the cryosar's row and column. Without individual resistors in series with each cryosar, however, a matrix of bistable cryosars could not be used in this fashion because any one cryosar that is in the low resistance "on" condition would tend to a.c. short circuit the row and column it is in. Of course, putting in individual resistors in series with each element would vitiate the advantage of the low-cost, simple fabrication indicated in Fig. 3.

THE COMPOUND CRYOSAR

Fortunately, series resistors can be built into the germanium wafer by modifying the impurity density and using the cryosar pre-breakdown resistance. The low-field pre-breakdown conductivity has been explained in terms of an impurity conduction[4] in which carriers migrate under the influence of an electric field by tunneling from impurity site to empty adjacent impurity site. Fig. 4 shows how this impurity conductivity varies with impurity concentration. Note that the conductivity can be varied over five orders of magnitude.

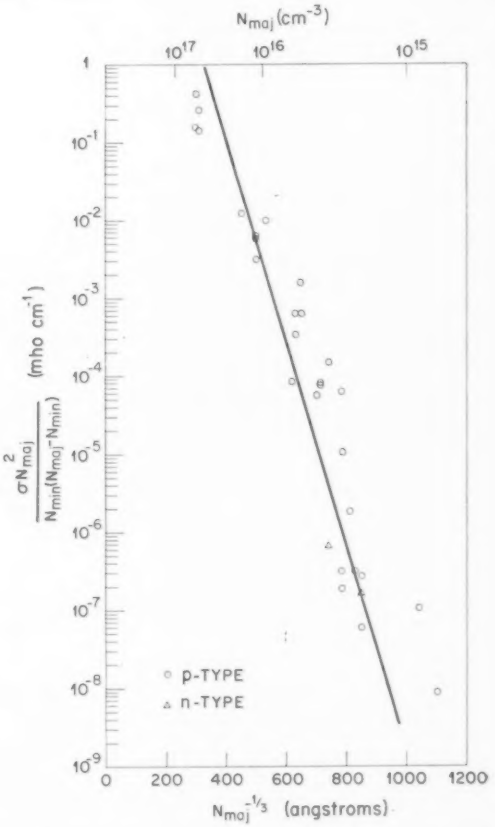


FIG. 4. Experimental values of impurity conductivity compared with a theoretical curve (see Ref. 2).

Referring to Fig. 1, in which the V–I characteristics of two cryosars are shown, the pre-breakdown conductivity of sample B is much larger than that of sample A because sample B has a higher majority impurity density. At higher fields samples with low impurity conductivity will exhibit non-ohmic behavior in the pre-breakdown region as in the case of sample A of Fig. 1. This may be due both to increasing numbers of carriers in the valence band and a changing mobility in this band.

The voltage–current characteristics of Fig. 1 are replotted on a linear scale in Fig. 5(a). Note

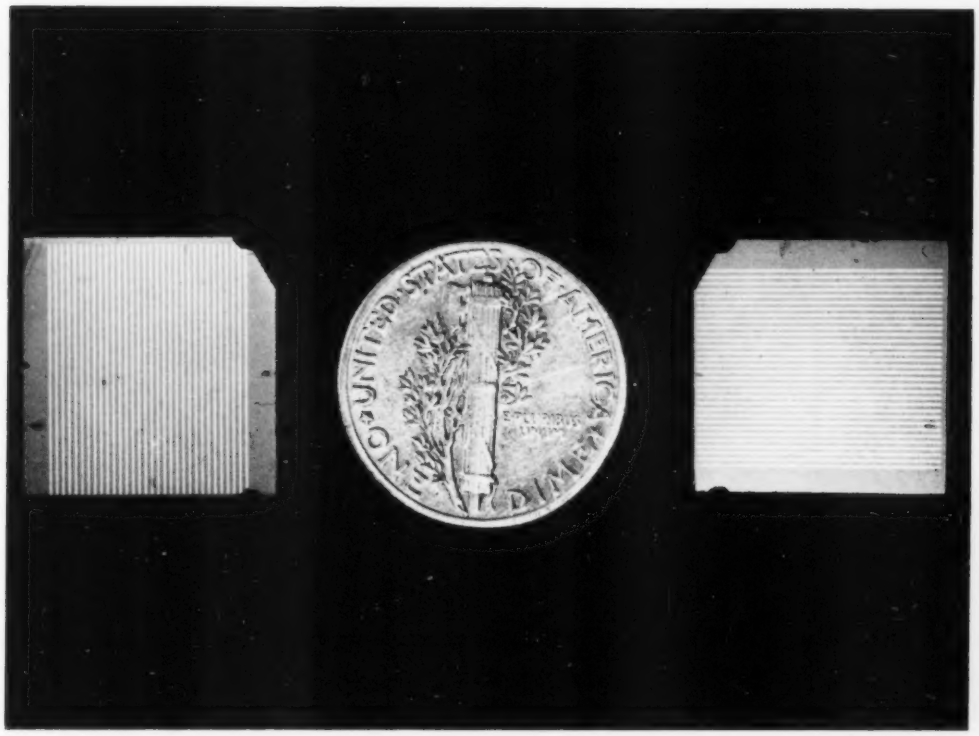


FIG. 3. Top and bottom views of a matrix of 1024 bistable cryosars. The wafer thickness is 0·005 in. The ohmic line contacts were made by evaporating and subsequently alloying aluminum.

[facing p. 102

the very low dynamic resistance when either of the elements are in their breakdown or "on" condition. This resistance is usually of the order of ohms. If samples *A* and *B* are placed in series, the resultant characteristic of the series configuration has three positive resistance regions and two negative resistance regions as shown in Fig. 5(b). The regions labeled I and II on the figure can be

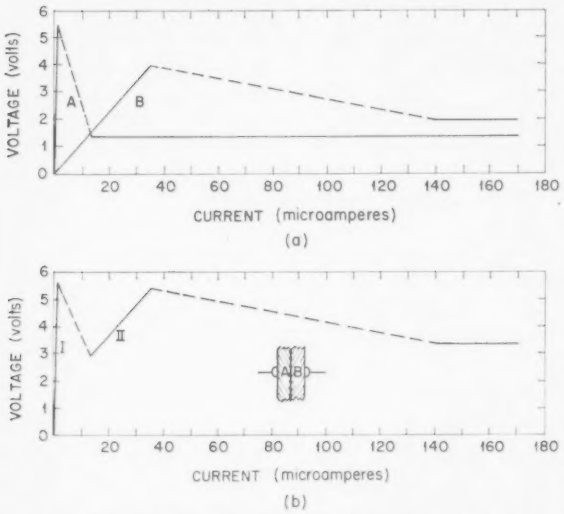


FIG. 5. (a) The voltage–current characteristics of cryosars *A* and *B* of Fig. 1, replotted on a linear scale. (b) The current–voltage characteristic expected for a compound cryosar consisting of two parts of equal thickness, one with the impurity density of *A*, the other with the impurity density of *B*.

used for the two stable states of this composite memory element; the resistance of these regions is high enough to eliminate the difficulties of shorting rows and columns in a coincident voltage memory as mentioned above.

Compound cryosars have been fabricated both by connecting two cryosars together with ohmic contacts and by using compensated crystals grown with the desired discontinuity.* When the grown compound-cryosar is in region II, the part of the wafer (*B*) that is not broken down conducts current in the impurity levels, while the other part of the wafer (*A*) conducts the current in the valence band (or conduction band for *n*-type). The voltage–current characteristic, however, does not seem to be affected by the junction between the two types of conduction. By adjusting the relative thickness of the two parts, the second voltage peak can be shifted vertically in voltage with respect to the first peak. By adjusting the impurity density of the more heavily doped part, the second voltage peak can be moved horizontally in current.

COINCIDENT-VOLTAGE MEMORY WITH NONDESTRUCTIVE READOUT

By making the second peak voltage less than the first, compound cryosars can be used as the elements in a coincident-voltage memory with nondestructive readout. The way in which they would be used can be explained in terms of the four load lines of Fig. 6. The intersection of the quiescent load line with the *V–I* characteristic determines the 0 and 1 stable states. Application of coincident *write* 1 voltage pulses shifts the load line to the *write* 1 position, leaving the cryosar in the 1 state when the pulses are removed. A single pulse is not large enough to switch the element from the 0 to 1 state. Coincident *write* 0 pulses set the element to the 0 state, a single pulse again not being large enough to switch the state. If the *read* pulse is applied with the cryosar in the 0 state, only a small increase in current will occur. But with the cryosar in the 1 state, the *read* pulse breaks down the heavily doped part, and the large resulting current indicates the 1 state. When the *read* pulse is removed, the cryosar is left in its initial state.

Compound-cryosar memory elements have been operated at a 10^{-7} sec cycle time using $1{\cdot}5\times10^{-8}$ sec pulses in essentially the way just described.[5] The elements have also been switched with pulses as short as 2×10^{-9} sec.

SPEED AND POWER CONSIDERATIONS

In order to switch a cryosar from a higher- to a lower-resistance region with a short pulse, the pulse amplitude must be larger than that calculated on the basis of the low-frequency peak voltage. For $1{\cdot}5\times10^{-8}$ sec pulses, the peak voltages are effectively increased by 20 per cent over their low-frequency values.[5] The time required to switch the cryosar to a higher-resistance region is extremely short and not usually measurable.

* Epitaxial growth from the vapor phase appears to be a very promising technique to fabricate compound cryosars with desired discontinuity in impurity concentration.

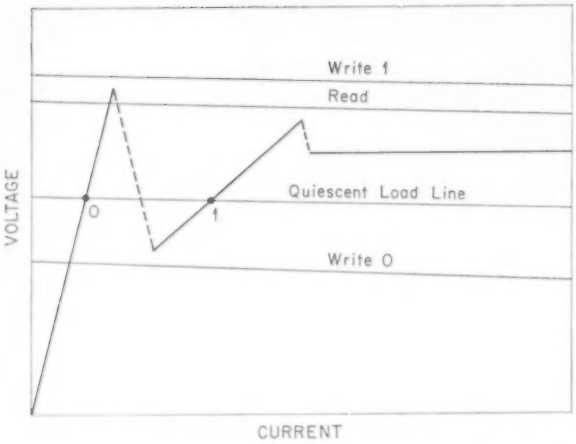


FIG. 6. Voltage–current characteristic of a compound cryosar with second peak voltage less than first peak voltage, with the four load lines indicating the manner in which this type of cryosar would be used as the element in a coincident-voltage memory with a non-destructive readout.

"Turn-off" times of the order of 10^{-10} sec are calculated.

In order to show that *RC* time constants would not limit the switching speed of a large-scale compound-cryosar memory, we may consider the following design. Using germanium wafers about 0·012 cm thick, about 0·001 cm would be for the part with the lower pre-breakdown resistance (part *B* in Fig. 5) and 0·011 cm would be for the part with the higher pre-breakdown resistance (part *A* in Fig. 5). Taking the width of the ohmic line contacts to be 0·01 cm, the impurity densities for part *B* to be $N_{maj} \simeq 7 \times 10^{15}$ cm^{-3} and $N_{min} \simeq 2 \times 10^{15}$ cm^{-3}, for part *A* to be $N_{maj} \simeq 1 \times 10^{15}$ cm^{-3} and $N_{min} \simeq 0{\cdot}5 \times 10^{15}$ cm^{-3}, and using the experimental data for these dopings, the compound cryosar would have the following electrical characteristics:

First peak voltage	$V_{p1} \sim 0{\cdot}30$ V
First peak current	$I_{p1} \sim 10^{-9}$ A
First valley voltage	$V_{v1} \sim 0{\cdot}15$ V
Second peak voltage	$V_{p2} \sim 0{\cdot}27$ V
Second peak current	$I_{p2} \sim 10^{-5}$ A
Second valley voltage	$V_{v2} \sim 0{\cdot}22$ V

The two stable states could then be 0·2 V at $<10^{-9}$ A (0) and 0·2 V at $\sim 4 \times 10^{-6}$ A (1). The pre-breakdown resistance of part *A* would be $\sim 1{\cdot}2 \times 10^4\ \Omega$.

When the compound cryosar is switched between its 0 state and its 1 state, it is part *A* which switches while part *B* remains in the pre-breakdown state, acting as a parallel combination of resistance and capacitance in series with part *A*. If one neglects the capacitance of part *B* and the resistance of part *A*, the worst possible case, the *RC* time constant for the cryosar described above is less than 10^{-9} sec.* The external circuitry necessary for the *write* and *read* operations will be shunted by the characteristic impedance of the row and column lines and their terminating resistors. With 50-Ω lines (a reasonable choice) up to 20 pF loading could be tolerated before the response time associated with the driving circuits is more than 10^{-9} sec.

In the above example the power dissipated by each compound cryosar in its 1 state is about 1μW, while the power dissipated in the 0 state is negligible. In *writing* 1, less than 0·6 mW of power is dissipated in the cryosar. In *reading* the state of a compound cryosar in the 1 state, less than 0·4 mW of power is dissipated in the cryosar. If we assume a duty cycle of one-fourth for either the *read* pulses or *write* pulses, and assume that not more than sixty-four cryosars will be read or written into at one time, an average of less than about 16 mW will be dissipated in the reading and writing operations. An additional average power of less than 30 mW will be dissipated because of the change in current and voltage during single *write* 1 and *write* 0 pulses. Thus a compound cryosar memory of 200,000 bits would dissipate less than 0·25 W. Further power dissipation will occur because of lead wires introduced into the liquid-helium bath and/or circuits placed in the

* This result is significantly different from that inferred by HUNTER[6]. In his treatment he did not take into account that the circuit capacitance could be isolated from the switching cryosar by a contiguous series resistor. Thus the value of capacitance directly across part *A* is an order of magnitude smaller than he assumed. Also, because the wafer thickness of matrices which would be used for high-speed, high-density memories is much less than that used in the preliminary experimental cryosars, the value of series resistance is considerably smaller than he assumed. His power dissipation considerations are similarly in error because he did not scale down the original experimental matrices to the proposed size.

bath to reduce the number of lead wires.* Experimental printed microstrip lines have been developed to reduce both the mass of the lead wires and the amount of heat transferred through them. Techniques of decoding in the bath are also under development.

CONCLUSIONS

A computer element operating at liquid-helium temperature has been described which can be used in large arrays for high-speed, high-density memories. One method of operating these elements in a coincident-voltage memory with nondestructive readout has been presented. The power dissipation of 200,000 elements is well within the range of a closed-cycle helium refrigerator with a modified 3 HP Freon compressor, similar to those used in large-window air conditioners.[7]

* By proper electrical design, it is possible to avoid d.c. power dissipation in the liquid-helium bath from the 50-Ω resistors which terminate the drive lines.

Acknowledgements—The authors wish to thank C. R. GRANT and I. MELNGAILIS for help in taking the data, W. H. LASWELL for fabricating the compound cryosars, and P. L. MOODY and A. E. PALADINO for growing the compensated germanium with desired impurity discontinuities. We are grateful to R. C. JOHNSTON for valuable discussions on computer circuitry.

REFERENCES

1. N. SCLAR and E. BURSTEIN, *J. Phys. Chem. Solids* **2,** 1 (1957); S. H. KOENIG and G. R. GUNTHER-MOHR, *Ibid.* **2,** 268 (1957).
2. A. L. MCWHORTER and R. H. REDIKER, *Proc. I.R.E.* **47,** 1207 (1959).
3. A. L. MCWHORTER and R. H. REDIKER, Paper presented at the *International Conference on Semiconductor Physics, Prague, Czechoslovakia,* 29 *August to* 2 *September* 1960.
4. E. M. CONWELL, *Phys. Rev.* **103,** 51 (1956); N. F. MOTT, *Canad. J. Phys.* **34,** 1357 (1956); T. KASUYA and S. KOIDE, *J. Phys. Soc. Japan* **13,** 1287 (1958).
5. R. C. JOHNSTON. Private communication.
6. L. P. HUNTER, *I.R.E. Trans. Electronic Computers* **8,** 503 (1959).
7. H. O. MCMAHON and W. E. GIFFORD, *Solid-State Electron.* **1,** 273 (1960).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 106–114. Printed in Great Britain

A POSSIBLE MECHANISM OF CRYSTAL GROWTH FROM THE MELT AND ITS APPLICATION TO THE PROBLEM OF ANOMALOUS SEGREGATION AT CRYSTAL FACETS

A. TRAINOR and B. E. BARTLETT

Research Laboratory, Mullard Southampton Works

(*Received* 1 *July* 1960; *in revised form* 16 *September* 1960)

Abstract—A possible mechanism of crystal growth from the melt is described by which stable facets are assumed to grow by nucleation at the point of critical supercooling, followed by rapid sheet growth across the facet. The velocity of this sheet growth would be independent of the rate of advance of the solid–liquid interface. The mechanism is used as the basis of a mathematical treatment of impurity segregation at a facet, taking into account the adsorption of impurity at the solid–liquid interface. An equation is derived which relates the apparent segregation coefficient at a facet to the velocity of sheet growth and the degree of impurity adsorption. It is shown that, for velocities in the expected range, the value of the segregation coefficient tends to a limiting value which may be greater than 1. The application of the theory to the related problem of dendrite growth is suggested.

Résumé—On décrit un mécanisme possible de développement de crystaux, en partant de la fonte. Par ce mécanisme, des facettes stables sont supposées se développer par nucléation au point de surfusion critique suivi d'un développement en feuille à travers la facette. La vitesse de ce développement en feuille serait indépendante du taux d'avance de l'interface solide–liquide. Ce mécanisme est employé comme base d'un traitement mathématique de séparation d'impuretés sur une facette en tenant compte de l'adsorption d'impuretés à l'interface solide–liquide. On dérive une équation qui relie le coefficient apparent de séparation sur une facette à la vitesse du developpement de la feuille et du degré d'adsorption d'impuretés. On démontre que pour des vitesses dans la gamme prévue, la valeur du coefficient de séparation tend à une limite qui serait plus grande que 1. On suggère l'application de cette théorie au problème apparenté de développement de dendrites.

Zusammenfassung—Zur Erklärung des Kristallwachstums aus der Schmelze wird angenommen, dass das Wachstum stabiler Kristallflächen mit der Keimbildung an einem Punkt kritischer Unterkühlung einsetzt, worauf ein schnelles Ebenenwachstum über die ganze Fläche erfolgt. Die Geschwindigkeit dieses Ebenenwachstums wäre von dem Fortschreiten der Grenzfläche zwischen Festkörper und Flüssigkeit unabhängig. Dieser Mechanismus dient als Grundlage für die mathematische Behandlung der Absonderung von Verunreinigungen auf der Kristallfläche, wobie die Adsorption der Verunreinigung an der fest–flüssig Grenzfläche berücksichtigt wird. Es lässt sich eine Gleichung ableiten, die eine Beziehung zwischen dem scheinbaren Absonderungskoeffizienten auf der Kristallfläche zu der Geschwindigkeit des Ebenenwachstums und dem Grad der Adsorption der Verunreinigung herstellt. Es wird gezeigt, dass für Geschwindigkeiten innerhalb des erwarteten Bereichs der Wert des Absonderungskoeffizienten einem Grenzwert zustrebt, der grösser als 1 sein könnte. Die entwickelte Theorie lässt sich auf das verwandte Problem des Dendritenwachstums anwenden.

NOTATION

σ no. atoms/unit area—subscripts as defined in text

ϕ arbitrary constant

$\bar{u}$ mean velocity—subscripts as defined in text

n no. atoms/unit volume—subscripts as defined in text

τ dwell time—subscripts as defined in text

D_i diffusion coefficient of the impurity within the adsorbed layer

v_y speed of sheet growth
β a function of growth speed as defined in the text
α two-dimensional segregation coefficient as defined in the text
ΔQ_a activation energy for the transfer of an atom from one site to an adjacent site on the solid surface.
Q adsorption energy
v_∞ maximum speed of growth of a solid sheet
k segregation coefficient—subscripts as defined in text

centration being often reasonably sharp. These and other related phenomena have been shown by HULME and MULLIN[2] in the case of indium antimonide, and by DIKHOFF[3] in the case of germanium, to be due to the presence of {111} facets in the growing interface, the segregation coefficient of the impurity being higher on the facet than off it. A particularly curious feature of this effect is that in the case of certain impurities the segregation coefficient of the impurity changes from a value less than 1·0 in regions off the facet to a value greater than 1·0 at the facet–liquid interface.

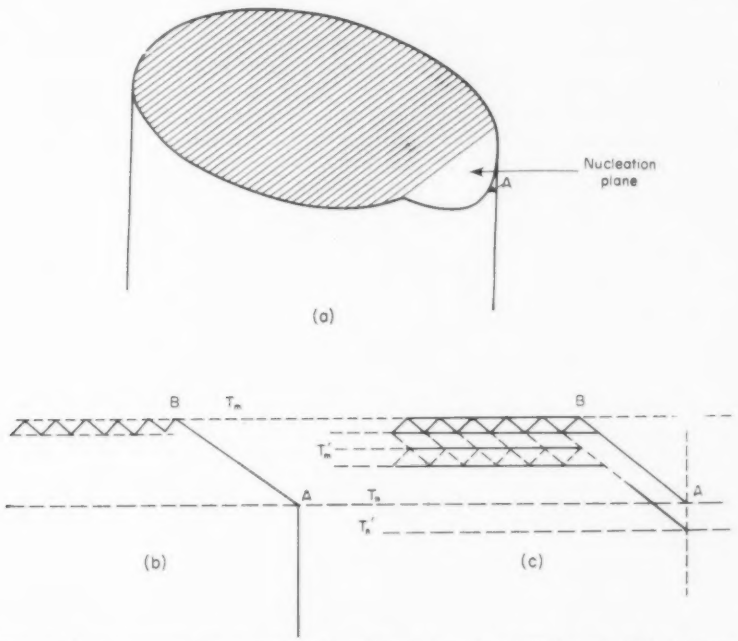


FIG. 1. [Step growth on a randomly oriented interface (T_m' and T_n' indicate previous position of T_m and T_n).

1. INTRODUCTION

SLICHTER and BURTON[1] observed that in twinned germanium crystals grown from a ⟨111⟩ seed the concentration of impurity differed on the two sides of the twin boundary by a considerable factor. They and others have also frequently observed the effect known as "coring" whereby the concentration of impurity in the central region of a crystal is considerably higher than that in other regions, the transition from higher to lower con-

Similar changes in the value of segregation coefficient have been observed by BENNETT and LONGINI[4] in the case of grown dendrites of germanium and in this laboratory by BARTLETT (unpublished) in the case of grown dendrites of indium antimonide. Such changes in the value of segregation coefficient cannot be explained on the basis of the theory of BURTON *et al.*[5] The treatment to be described is related only to the normal crystal-growing problem, though a

suggestion of its application to dendritic growth is outlined.

2. PROPOSED MECHANISM OF CRYSTAL GROWTH

Experimental difficulties in examining crystal–liquid interfaces in the case of crystals grown from the melt have combined with the difficulties experienced in the theoretical treatment of the liquid state to retard progress in the understanding of the mechanisms by which such crystals grow. The growth of silicon and germanium crystals apparently free of dislocations tends to suggest that a FRANK[6–8] mechanism is excluded, though the absence of screw dislocations has not yet been conclusively proved. It is therefore profitable to consider other possible mechanisms.

Crystals will always tend to grow towards a shape bounded by the slowest-growing planes, but, in the case of crystal growth from the melt as normally carried out, crystal growth is forced by the imposition of a moving temperature gradient. Under these conditions the crystal is prevented from assuming its equilibrium shape. Fig. 1 illustrates a possible configuration for a randomly oriented solid–liquid interface in a material having the diamond or zinc blende lattice. The most close-packed, and therefore the slowest-growing, plane in such a lattice is the {111} plane. If all growth occurs in a direction parallel to this plane the interface will consist of a series of steps as shown, probably one atom high. The isotherm T_m limits the growth of the sheets and it will be readily seen that the step configuration is a stable one, the steps moving across the face of the growing crystal as the isotherm moves normal to the face [see Fig. 1(c)]. The initiation of new steps is, however, necessary and this requires the nucleation of a (111) plane. The addition of atoms to a step is a process more readily accomplished than the formation of a new nucleus, and some degree of supercooling is almost certainly necessary to initiate this process. The point A, therefore, on the generating plane AB will lie on an isothermal T_n, where $(T_m - T_n)$ is the degree of supercooling required to nucleate the plane. Once the nucleus has formed, a monatomic sheet will grow at a rate independent of the speed of movement of the isotherms and dependent only on such considerations as the maximum rate of incorporation of liquid atoms into the growing solid sheet, and the rate of removal of latent heat. This fast sheet growth will continue until the growing sheet meets the isotherm T_m, when another step results.

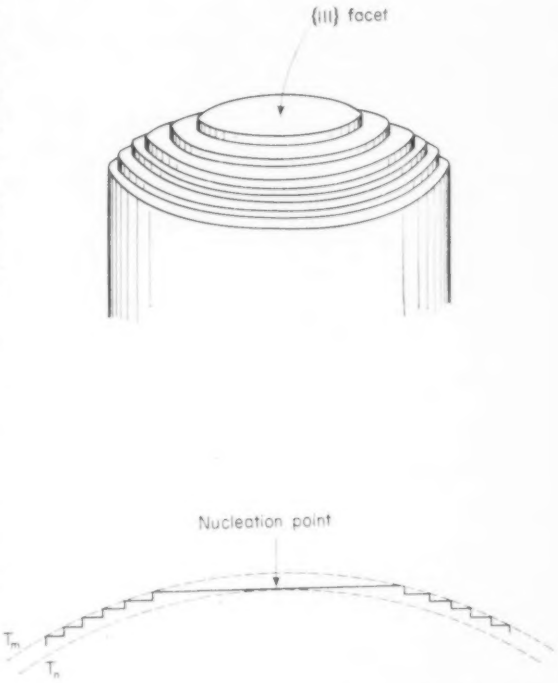


FIG. 2. Schematic representation of convex interface of a crystal grown in the ⟨111⟩ direction (roughly concentric isotherms).

Fig. 2 illustrates the similar conditions which apply for a crystal grown in the ⟨111⟩ direction with non-linear isotherms. A {111} facet can be formed and this cannot be nucleated until some point on the facet reaches the isotherm T_n. Rapid sheet growth then occurs maintaining a stable facet in the interface. Growth off the facet is again by means of steps and no nucleation is required. The growth speed off the facet is thus controlled by the rate of movement of the isotherm system.

3. APPLICATION OF THE PROPOSED MECHANISM TO THE PROBLEM OF ANOMALOUS SEGREGATION

On consideration of the proposed mechanism of

crystal growth an immediate difference in the conditions existing on a growing {111} facet from those on the rest of the interface becomes apparent, namely that the speed of growth of the monatomic sheets is likely to be very much faster on the facet than off it.

It is possible that the mechanism of the segregation of impurities from a growing crystal may have as its first step the adsorption of the impurity and the host atoms at the surface of the solid, leading to an interfacial impurity concentration differing from the concentration in the bulk of the melt. The combined effect of velocity of sheet growth and of impurity and host-atom adsorption will now be considered.

4. THE EFFECT OF ADSORPTION

It can readily be shown that the surface concentrations of solvent and solute atoms (or molecules) at a solid surface in contact with a solution are given by

$$\sigma_i = \phi \bar{u}_i n_i (1 - \sigma_s/\sigma_0 - \sigma_i/\sigma_0)\tau_i$$

$$\sigma_s = \phi \bar{u}_s n_s (1 - \sigma_s/\sigma_0 - \sigma_i/\sigma_0)\tau_s$$

where σ = surface concentration (atoms/cm^2), $\bar{u}$ = mean particle velocity in solution, n = no. of atoms (molecules)/cm^3 in solution, σ_0 = no. of sites/cm^2 solid surface, τ = mean dwell time of an atom (molecule) at the surface, the subscripts i and s refer to solute and solvent respectively, and ϕ = constant.

These equations refer to an equilibrium state, but it can readily be shown that the effect of growth velocity normal to the solid interface is negligible for low growth velocities.

From this

$$\sigma_i/\sigma_s = \bar{u}_i n_i \tau_i / \bar{u}_s n_s \tau_s \qquad (1)$$

We can now regard the adsorbed layer as a two-dimensional liquid and consider the growth of a monatomic sheet through this liquid. The total number of atoms/cm^2 in the two-dimensional liquid is seen to be $(\sigma_i + \sigma_s)$. In the grown sheet, however, the number of atoms/cm^2 = σ_0. Thus there will be an area change and this must be taken into account.

If we define a term α as a two-dimensional segregation coefficient applying at the interface between the growing solid sheet and the adsorbed layer, then

$$\alpha = \sigma_{ib}/\sigma_i\{\sigma_0/(\sigma_i + \sigma_s)\}$$

where σ_{ib} is the concentration of impurity in the solid sheet (atoms/cm^2) and the bracketed term is a correction term for the area change already mentioned.

Then

$$\alpha = \sigma_{ib}\sigma_s/\sigma_i\sigma_0$$

for dilute solution where $\sigma_i \leqslant \sigma_s$.

This gives rise to a solid concentration

$$C_s = \sigma_{ib} N^{1/3}$$

(where N = no. atoms/cm^3 in solid), and, since liquid concentration = $C_L = n_i$,

$$k = C_s/C_L = \sigma_{ib} N^{1/3}/n_i = \alpha(\sigma_i/\sigma_s)(\sigma_0 N^{1/3}/n_i)$$

Under infinitely slow growth conditions both normal and parallel to the plane of the interface we can write

$$k_0 = \alpha_0(\sigma_i/\sigma_s)\cdot(\sigma_0 N^{1/3}/n_i) \qquad (2)$$

where k_0 = quasi-equilibrium segregation coefficient,

$$\sigma_0 \simeq N^{2/3} \text{ and } n_s \simeq N$$

We can therefore write from (1) and (2)

$$k_0 = \alpha_0 \bar{u}_i \tau_i / \bar{u}_s \tau_s$$

The determination of $\bar{u}_i$ and $\bar{u}_s$ will be discussed in a later section, when it will be seen that $\bar{u}_i$ and $\bar{u}_s$ should not be markedly different and we may thus write

$$k = \alpha \tau_i / \tau_s \qquad (3a)$$

$$k_0 = \alpha \tau_i / \tau_s \qquad (3b)$$

5. THE EFFECT OF THE VELOCITY OF SHEET GROWTH

The picture with which we are now concerned is shown diagrammatically in Fig. 3. At any point sufficiently distant from the growing sheet front at $y = 0$ $(y \to \infty)$ there will be no concentration gradient of impurity within the sheet and a dynamic balance exists between the number of atoms being adsorbed from solution and the number desorbed in unit time. A finite sheet-growth velocity v_y is equivalent to a flow within the two-dimensional liquid equal to $-v_y$. Segregation of impurity at the solid/adsorbed-layer interface with a segregation coefficient $\alpha_0 \neq 1$ will give rise to a diffusion

gradient in front of the growth front. This diffusion gradient will lead to an unbalance in the number of atoms being adsorbed and desorbed in unit time. Let us assume that α_0 is always less than 1. Then more atoms will be desorbed than adsorbed at any finite y, leading to an impurity current from the adsorbed layer into the melt. Under conditions of dynamic equilibrium there will be, therefore, a balance of impurity currents in an element dy such that the nett change in impurity content of the element due to the effect of concentration gradient, of liquid flow $(-v_y)$ and of loss to the bulk of the solution, is zero.

The equation describing this system is

$$D_i \partial^2\sigma/\partial y^2 + v_y \partial\sigma/\partial y - [(1/\tau_i + \phi\tilde{u}_i n_i/\sigma_0)\sigma - \phi\tilde{u}_i n_i(1-\sigma_s/\sigma_0)] = = \partial c/\partial t = 0$$

The first term of the equation describes the impurity flux due to the concentration gradient, D_i being the diffusion coefficient of impurity in the adsorbed layer. The second term describes the contribution of the growth velocity to the impurity flux, and the bracketed term allows for the transport of impurity from the adsorbed layer to the liquid.

To a first approximation the growth velocity in the x-direction ("pulling speed") will govern only the rate at which sheets are nucleated, and will have no effect on the concentration and behaviour of the impurity within the growing sheet for normal values of v_x (0–100 cm/hr).

Boundary conditions are

$$\sigma = \sigma_i \qquad y = \infty$$

$$D_i \partial\sigma/\partial y + v_y(\sigma^* - \sigma_{ib}) = 0 \qquad y = 0$$

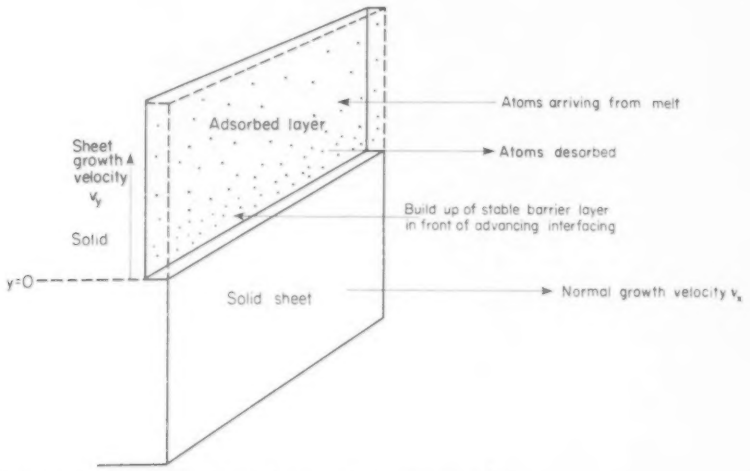


FIG. 3. Schematic representation of the growth system on a facet.

where σ^* = sheet concentration in the adsorbed layer at the growth front.

The first boundary condition is given by the equilibrium velocity-independent adsorbed-layer concentration in regions remote from the growth front. The second by the balance of fluxes at the growth front necessary for the establishment of a stable concentration gradient. It is similar to the BURTON *et al.*[5] boundary condition.

It is convenient to write

$$\left.\begin{aligned} A &= 1/\tau_i + \phi\tilde{u}_i n_i/\sigma_0 \\ B &= \phi\tilde{u}_i n_i(1-\sigma_s/\sigma_0) \end{aligned}\right\} \qquad (4)$$

This solves to

$$\sigma = \sigma_i + [v(\sigma^* - \sigma_{ib})/\beta D_i]\exp(-\beta y)$$

$$\beta = [v + \sqrt{(v^2 + 4AD_i)}]/2D_i \qquad (5)$$

for $y = 0$; therefore

$$\sigma^* = \sigma_i + v(\sigma^* - \sigma_{ib})/\beta D_i$$

Putting

$$\alpha_0 = \sigma_{ib}/\sigma^* \text{ and } \alpha = \sigma_{ib}/\sigma_i$$

this reduces to

$$\alpha = \alpha_0/[1 - v/\beta D_i(1-\alpha_0)]$$

From (3a), (3b) and (5)

$$k = k_0/[1 - v/\beta D_i(1 - k_0\tau_s/\tau_i)] \qquad (6)$$

Equation (6) compares the segregation of the impurity on the crystal facet to the normal segregation occurring in randomly oriented regions of the solid–liquid interface, since k is the segregation coefficient for regions where fast sheet growth is occurring and k_0 the equilibrium segregation coefficient to be expected in regions where the sheet growth is slow and controlled by the moving temperature gradient.

In order to use the equation it is necessary to estimate some of the constants involved, since several of them are unknown.

6. VALUES OF THE CONSTANTS

Mean velocities u_i and u_s

In the case of dilute solution the movement of a solute atom is governed largely by the rate of arrival of adjacent holes of sufficient size to allow the atom to exchange its former site for an adjacent one. Provided that the solvent atom or molecule does not differ too greatly in size from the solute atom the mean velocities of the two species will be very nearly identical: $\bar{u}_i = \bar{u}_s$.

Values of mean velocity in liquids are known to be of the order 10^3–10^4 cm/sec.

The surface diffusion coefficient D_i

The adsorbed layer at the surface of the solid has been regarded as a two-dimensional liquid, but consideration of the total number of adsorbed atoms/cm^2 suggests that its condition is nearer to that of a dense gas. Thus the degree of occupation will not be unduly high. The adsorbed atoms are fairly tightly bound to the solid surface and intuitively a diffusion coefficient between the values for the liquid and the solid would be expected, giving a range from 10^{-9} cm^2/sec to 10^{-4} cm^2/sec.

Since we are dealing with the surface of a crystalline solid we may assume a regular arrangement of preferred adsorption sites corresponding to the regular atomic spacing. If an adsorbed atom is to diffuse over the surface, therefore, it must do so by surmounting the energy barrier between one preferred site and a neighbouring site. This leads to the concept of the "hopping molecule" (see DE BOER[9]). The halting time between hops has been shown to be given by

$$\tau' = \tau'_0 \exp(\Delta Q_a/RT)$$

where ΔQ_a is the activation energy for transfer to another site and τ'_0 is of the order 10^{-13} sec. ΔQ_a would be expected to be between one half and two thirds of the adsorption energy.

If the transit time of the atom from one site to the next is negligible compared with its dwell time between hops (τ') it is possible to calculate the diffusion coefficient D_i. KRUYER[10] has shown this to be given by:

$$D_i = a^2/4\tau'$$

where a = hopping distance [distance between atoms in the (111) plane in this case].
Taking ΔQ_a in the range 5–25 kcal/mole

	τ' (sec)	D_i (cm^2/sec)
InSb	4×10^{2} to $1{\cdot}5\times10^{-10}$	6×10^{-18} to $1{\cdot}7\times10^{-5}$
Ge	2×10^{-9} to 10^{-12}	10^{-7} to 2×10^{-4}
Si	4×10^{-12} to 2×10^{-13}	2×10^{-4} to $4{\cdot}5\times10^{-3}$

The calculated values of D_i lie within the expected range with the exception of the value of 6×10^{-18} for indium antimonide, which corresponds to an assumed value of 25 kcal/mole for ΔQ_a. This value of ΔQ_a is undoubtedly excessive in the case of indium antimonide and the value calculated on the basis of a barrier height of 5 kcal/mole is more realistic.

Dwell times τ_s and τ_i

The dwell time of an adsorbed atom at a solid surface in contact with a solution is given by

$$\tau = \tau_0 \exp(Q/RT)$$

$$\tau_0 = 4{\cdot}75\times10^{-13}\sqrt{(MV^{2/3}/T_s)}$$

where Q = adsorption energy, τ_0 = time of

vibration of the lattice, M = molecular (or atomic) weight, V = molar volume, T_s = melting point (°K) and τ_0 is usually of the order 10^{-13}.

If we take Q to be in the range 20–60 kcal/mole (i.e. the order of the single-bond energies for most diamond and zinc blende lattices), τ lies between 10^{-7} and 10^{-2} sec for germanium.

In the case of substitutional impurities the ionic charge will contribute to the adsorption energy.

7. DISCUSSION

In interpreting equation (6) it is useful to use the dimensionless number $X = v(D_i A)^{1/2}$, when one graph may be used to determine the behaviour of the function over a wide range of conditions. Thus from (5)

$$v/\beta D_i = 2v/[v+\sqrt{(v^2+4D_iA)}]$$
$$= 2X/[X+\sqrt{(X^2+4)}]$$

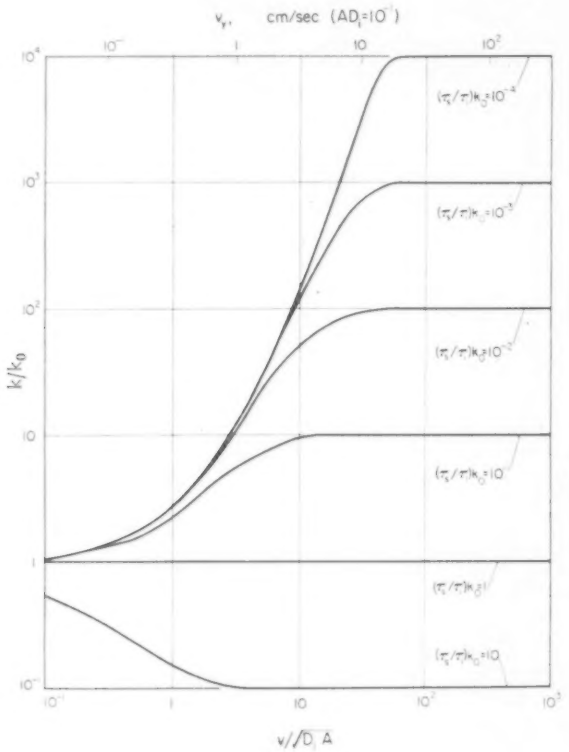


FIG. 4.

Thus τ_i/τ_s could be greater than 1, and a difference in behaviour between n- and p-type impurities would be expected.

The value of τ_i for a given value of Q is strongly dependent on the melting point of the solvent, τ_i decreasing with increasing temperature. This is in agreement with observation, since the anomalous segregation effect appears to be most marked in InSb and of least significance in the case of silicon for the materials investigated.

Fig. 4 shows the result of plotting k/k_0 against X for various values of $\tau_s k_0/\tau_i$.

It will be seen that for values of $\tau_s k_0/\tau_i < 1$ the ratio of the segregation coefficient on the facet (k) to the normal segregation coefficient off the facet (k_0) rises above 1·0 and approaches a limiting value of $\tau_i/k_0\tau_s$.

For high values of X, therefore, the segregation coefficient observed will differ from the equilibrium segregation coefficient provided that $\tau_s k_0/\tau_i$ is not

$=1$. This implies that if the velocity of sheet growth is high, as is expected to be the case on a facet in the growing interface, the segregation coefficient will depend only on the degree of impurity adsorption at the interface.

The top scale in Fig. 4 gives actual values of sheet velocity assuming reasonable values of D_i and τ_i.

Certain other conclusions may be drawn from Fig. 4.

(1) Since τ_i/τ_s is a measure of the adsorption of the impurity at the solid–liquid interface, it can be seen that the impurity need not be more strongly adsorbed than the host atom for the segregation coefficient at the crystal facet to be larger than the value off the facet, as it is only necessary that $\tau_s k_0/\tau_i$ be $<1{\cdot}0$. However, the increase in segregation coefficient at the facet will be more marked the stronger is the impurity adsorption, and the latter must be greater than that of the host atom if the value of k is to change from a value less than 1 to one greater than 1.

(2) If an effect is to be observable, the velocity of sheet growth would have to be greater than some critical value in the range 10^{-1}–10^3cm/sec, dependent on the values of D_i and A.

The speed of growth of a straight step has been calculated for crystals grown from the vapour (see VERMA[11]) and is given by

$$v_\infty = 2(\alpha-1)X_s\nu \exp(-W_s/kT)$$

where $(\alpha-1)$ = supersaturation of the vapour, X_s = mean distance of diffusion of an adsorbed molecule, ν = frequency factor of the order of the atomic frequency of vibration and W_s = energy of evaporation from kinks in the growing step on to the surface $\simeq W/2$ where W = the evaporation energy. Calculation of v_∞ for growth under typical conditions gives

$$v_\infty \simeq 10^2\text{–}10^3 \text{ cm/sec}$$

The rate of growth of a step from the melt would be expected to be at least of the same order, and this growth velocity is of the right order to produce the conditions for anomalous segregation.

8. DENDRITIC GROWTH

An anomalous segregation behaviour has been observed in dendritic crystals of germanium by BENNETT and LONGINI[4] which seems to exhibit features in common with the behaviour on crystal facets. In some cases, the sign of $(1-k)$ changes in the case of dendrites grown at high growth rates. Similar effects have been observed in indium antimonide by BARTLETT who has measured the effect of dendritic growth over the range 1–20 cm/min. BENNETT and LONGINI have suggested a mechanism of dendritic growth involving the formation of a longitudinal twin as a nucleation centre. The growth of single-crystal whiskers[12] has been shown to take place on a thin filamentary core which extends a considerable distance in front of the growing whisker, the whisker being built up around the core at a slower rate. It may be, therefore, that dendritic growth proceeds in a somewhat similar manner, a very thin platelet preceding the main growth front, the build-up of the bulk of the dendrite being by sheet growth on the {111} planes parallel to the platelet, as has been suggested by BILLIG[13] and by BENNETT and LONGINI[4]. If this were the case a similar mechanism to the one proposed could apply. The velocities of dendritic growth, 1–20 cm/min (10^{-2}–1 cm/sec), are of the right order for the effect to be observable. The results will be complicated, however, by the presence of a second effect due to a mechanism of the same type as the one described by BURTON *et al.*[5], but values of segregation coefficient greater than 1 may be due to a mechanism similar to that described.

It is of interest that, using the expected values for the constants involved, the term A is concentration dependent. Since velocities in sheet growth are high, the values of k encountered should have levelled off at the upper limit and therefore would not be expected to show much concentration dependence. In the case of dendrites, however, where growth velocities are lower, the effect should be apparent in certain cases, and has been observed by BENNETT and LONGINI.

For example:

$$\tau_i = 10^{-4}\,\text{sec},\ D_i = 10^{-5}\,\text{cm}^2\,\text{sec}^{-1},\ \phi\tilde{u} = 10^3\text{cm/sec}$$

$$\sigma_0 \simeq 10^{15}$$

Then

$$A = 4D_i(1/\tau_i + \phi\tilde{u}n_i/\sigma_0) = 4\times10^{-5}(10^4 + 10^{-12}.n_i)$$

Thus, for low values of n_i, A is independent of

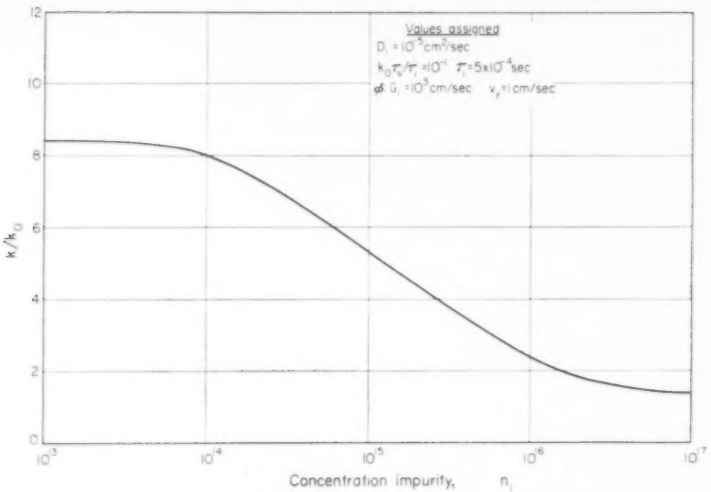

Values assigned
$D_i = 10^{-5}$ cm²/sec
$k_0\tau_s/\tau_i = 10^{-1}$ $\tau_i = 5\times10^{-6}$ sec
ϕ $\bar{u}_i = 10^3$ cm/sec $v_y = 1$ cm/sec
k/k0
Concentration impurity, n_i

FIG. 5. Variation of segregation coefficient with melt concentration for fixed growth velocity (v_y).

concentration but for values $n_i < 10^{14}$ concentration becomes increasingly important (Fig. 5).

9. CONCLUSION

A mechanism for crystal growth from the melt has been suggested by which all growth occurs in thin sheets, probably monatomic. The surface of these sheets is a {111} plane and growth is in a direction perpendicular to the ⟨111⟩ direction. The speed of growth of such sheets has been shown to be controlled by the movement of the melting-point isotherm except in cases where {111} facets are formed in the solid–liquid interface, when growth speed would reach speeds governed only by considerations such as the rate of arrival of solvent atoms at the growth step. This mechanism has been used as a basis for the calculation of the effect of impurity and host-atom adsorption combined with sheet growth on the segregation of impurities at a crystal facet. It has been shown that, using values of sheet velocity and heats of adsorption not inconsistent with those expected, the anomalous segregation observed in diamond and zinc blende lattice crystals could be explained.

Some observations have been made on the anomalous segregation behaviour of certain impurities in dendritic crystals grown from the melt.

Acknowledgements—The authors wish to thank the directors of Mullard Limited for permission to publish this paper. We also wish to thank Dr. J. B. MULLIN and Dr. K. F. HULME of RRE, Malvern, and Dr. M. SMOLLETT and other members of these laboratories for helpful discussions.

REFERENCES

1. W. P. SLICHTER and J. A. BURTON, *Transistor Technology* Vol. 1, Chap. 6. Van Nostrand, New York (1958).
2. K. F. HULME and J. B. MULLIN, *Phil. Mag.* **4,** 1286 (1959).
3. J. A. M. DIKHOFF, *Solid-State Electron.* **1,** 202 (1960).
4. A. I. BENNETT and R. L. LONGINI, *Phys. Rev.* **116,** 53 (1959).
5. J. A. BURTON, R. C. PRIM and W. P. SLICHTER, *J. Chem. Phys.* **21,** 1987 (1953).
6. F. C. FRANK, *Disc. Faraday Soc.* No. 5, 48 (1949).
7. F. C. FRANK, *Proc. Roy. Soc.* A**201,** 586 (1950).
8. F. C. FRANK, *Phil. Mag.* **41,** 200 (1950).
9. J. B. DE BOER, *The Dynamical Character of Adsorption.* The Clarendon Press, Oxford (1953).
10. S. KRUYER, *Proc. Acad. Sci. Amst.* B**56,** 274 (1953); see also Ref. 9.
11. A. R. VERMA, *Crystal Growth and Dislocations.* Butterworths, London (1953).
12. J. E. GORDON, *New Scientist* **6,** 986 (1959).
13. E. BILLIG, *Proc. Roy. Soc.* A**229,** 347 (1955).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 115–122. Printed in Great Britain

AN ANALYSIS OF THE PERFORMANCE OF THERMOELECTRIC DEVICES MADE FROM LONG-LIFETIME SEMICONDUCTORS

R. N. HALL

General Electric Research Laboratory, Schenectady, New York

(*Received* 13 *October* 1960; *in revised form* 26 *October* 1960)

Abstract—The performance of an ohmic–*n*–*p*–ohmic junction rectifier made from a long-lifetime semiconductor is analyzed as a thermoelectric generator and as a refrigerator. An upper limit to the efficiency of this device is shown to be inferior to that of a conventional (zero-lifetime) thermoelement.

Arguments are presented to show that any semiconductor device, in which non-equilibrium concentrations of minority carriers play a part in determining the electrical characteristics, will have a poor thermoelectric efficiency. It seems safe to make the generalization that lifetime is not an important material property for semiconductor thermoelements.

Résumé—Le rendement d'un redresseur à jonction *p–n* ohmique utilisant un semiconducteur à longue durée de vie est analysé en tant que générateur thermoélectrique ainsi qu'en tant que réfrigérant. La valeur maximum du rendement de ce produit est inférieure à celle des éléments thermoélectriques habituels.

Des arguments sont présentés pour démontrer que n'importe quel semiconducteur, dans lequel les concentrations inéquilibrées de porteurs minoritaires jouent un rôle pour déterminer les caractéristiques électriques, aura un faible rendement thermoélectrique. Il semble permis d'émettre la généralisation suivante: la longévité n'est pas une propriété matérielle importante pour les semiconducteurs thermoélectriques.

Zusammenfassung—Der Wirkungsgrad eines Gleichrichters mit *n–p*-Übergang zwischen zwei ohmschen Kontakten, der aus einem langlebigen Halbleiter hergestellt wurde, wird bie Verwendung als thermoelektrischer Generator und als Kühler analysiert. Die obere Grenze des Wirkungsgrads dieses Geräts ist dem eines üblichen Thermoelements (Lebensdauer Null) unterlegen.

Es wird nachgewiesen, dass jedes Halbleitergerät, in dem, wenn kein Gleichgewichtszustand existiert, die Dichte der Minoritätsträger die elektrischen Charakteristiken mitbestimmt, einen schlechten thermoelektrischen Wirkungsgrad besitzen muss. Man kann wohl ganz allgemein feststellen, dass die Lebensdauer für Halbleiter-Thermoelemente keine massgebende Materialeigenschaft ist.

1. INTRODUCTION AND SUMMARY

CONVENTIONAL thermocouples are constructed from semiconductors having lifetimes which are short enough that the carrier concentrations may be assumed to have the values which correspond to equilibrium at the local temperature. This paper investigates the possibility of obtaining higher efficiencies through the use of long-lifetime materials. Attention is directed primarily toward the thermoelectric behavior of semiconductor structures containing *p–n* junctions. The thermoelectric properties of uniformly doped semiconductors with long carrier lifetimes have been discussed by TAUC[1].

The analysis makes use of a formulation which is commonly employed in the treatment of rectifier and transistor characteristics. A number of simplifying approximations are made in order to obtain results which are not unduly cumbersome. However, it is believed that the results are suffici-

ently accurate to provide a reliable comparison between the types of thermoelectric generators that are investigated.

In the first sections of the paper, the electrical characteristics of p–n junctions containing thermal gradients are calculated. These results are then used to obtain the performance of the device as a power generator and as a thermoelectric refrigerator. The results of an analysis of an n–i–n type of structure are also presented. Comparisons with the performance of conventional short-lifetime thermocouples indicate that carrier lifetime is not a material parameter than can be increased to advantage for thermoelectric purposes. While it is difficult to prove that increases in efficiency over that of a conventional thermocouple cannot be obtained with *any* kind of long-lifetime thermoelectric device, physical arguments are presented to show that this is probably the case.

2. DEFINITION OF SYMBOLS

a, b coefficients defined by equations (9) and (10) (cm^{-3})
D diffusion coefficient ($cm^2\ sec^{-1}$)
e base of natural logarithms
E_g energy gap (V)
G conductance ($\Omega^{-1}\ cm^{-2}$)
J current density ($A\ cm^{-2}$)
k Boltzmann's constant (energy deg^{-1})
m coefficient describing electron scattering
N_c density of states in conduction or valence band (cm^{-3})
n, p electron and hole concentrations (cm^{-3})
P power generated ($W\ cm^{-2}$)
Q heat transfer ($W\ cm^{-2}$)
q electron charge (coulombs)
T absolute temperature (deg)
V applied voltage (V)
z thermoelectric figure of merit (deg^{-1})
α coefficient defined by equation (13)
η thermoelectric efficiency, ratio of electrical power generated to thermal power supplied
κ thermal conductivity ($W\ cm^{-1}\ deg^{-1}$)
ϕ internal junction potential (V)

3. ANALYSIS

A. *Isothermal junction*, $\tau = \infty$

Consider, first, a p–n junction at a uniform temperature, with ohmic contacts (infinite recombination rate) at $x = -l$ and $x = +l$. The junction is located at $x = 0$, as shown in Fig. 1. We assume equal electron and hole mobilities, with l small compared to a diffusion length, i.e. $\tau = \infty$. The n- and p-type regions are assumed to be uniform except within a distance δ of the junction, where $\delta \ll l$.

With no applied voltage, the net current through the junction is zero. However, electrons and holes are actually flowing back and forth across the junction, and it is instructive to calculate the magnitude of each component. The electron flow can be obtained from the superposition of the solutions of two hypothetical situations: (1) The above junction with ohmic contact at $x = -l$, but with boundary condition $n = 0$ at $x = l$, and (2) the same, but with the ohmic contact at $x = l$ and boundary condition $n = 0$ at $x = -l$.

The first situation gives the flow of electrons from left to right across the junction. Electrons in concentration n_n on the n-type side of the junction give rise to a concentration $n_n \exp(-q\phi_0/kT)$ on the p-type side. These diffuse to the contact at $x = l$, giving rise to a current due to electrons flowing up over the potential hill at this junction,

$$J_{e.\,\mathrm{right}} = \frac{qDn_n}{l} \exp(-q\phi_0/kT) \qquad (1)$$

(We adopt the convention that a current of either electrons or holes flowing in the forward direction has a positive sign. Similarly for applied voltages.)

The second situation gives the flow due to electrons generated at $x = l$ diffusing to the junction and falling down the potential hill. This component of the electron current is

$$J_{e.\,\mathrm{left}} = -\frac{qDn_p}{l} \qquad (2)$$

Since $n_p = n_n \exp(q\phi_0/kT)$, the flow of electrons in each direction across the junction is the same, in the absence of any applied voltage. If a forward voltage V is applied across the junction, the exponential in (1) is replaced by

$$\exp[-q(\phi_0 - V)/kT]$$

The current due to electrons can be put in the familiar form:

$$J_e = \frac{qDn_p}{l}[\exp(qV/kT) - 1] \qquad (3)$$

Similar equations describe the flow of holes.

In order to enter the semiconductor at $x = -l$, electrons must climb a hill of height ϕ_n where

$$\phi_n = \frac{kT}{q} \ln\left(\frac{N_c}{n}\right) \tag{4}$$

N_c is the effective density of states in the conduction band.

This equation assumes a non-degenerate semiconductor ($n \ll N_c$), and neglects a term of magnitude mkT/q which represents the kinetic energy of an electron in the conduction band. m is usually between $+2$ and $+4$,[2] depending upon the nature of the electron-scattering processes. This term must be included in the analysis of conventional thermocouples, where the thermoelectric power may be quite small. However, if the presence of the minority carriers is to contribute to the thermoelectric performance by virtue of the assumed long lifetime, the junction must be constructed from less heavily doped semiconductor material, and under these circumstances this term becomes of minor importance.

An electron which passes through the rectifier in the forward direction extracts amounts of energy $q\phi_n$ at $x = -l$, and $q(\phi_0 - V)$ at $x = 0$, and gives up an amount $q(E_g - \phi_p)$ at $x = l$. Holes behave similarly, so that a flow of current through the rectifier in the forward direction tends to cool the junction by removing heat at $x = 0$ and adding a somewhat greater amount at the two ohmic contacts. Joule heating may be neglected since in the above structure the "IR drop" in the end regions is assumed to be small compared with the potential hill at $x = 0$.

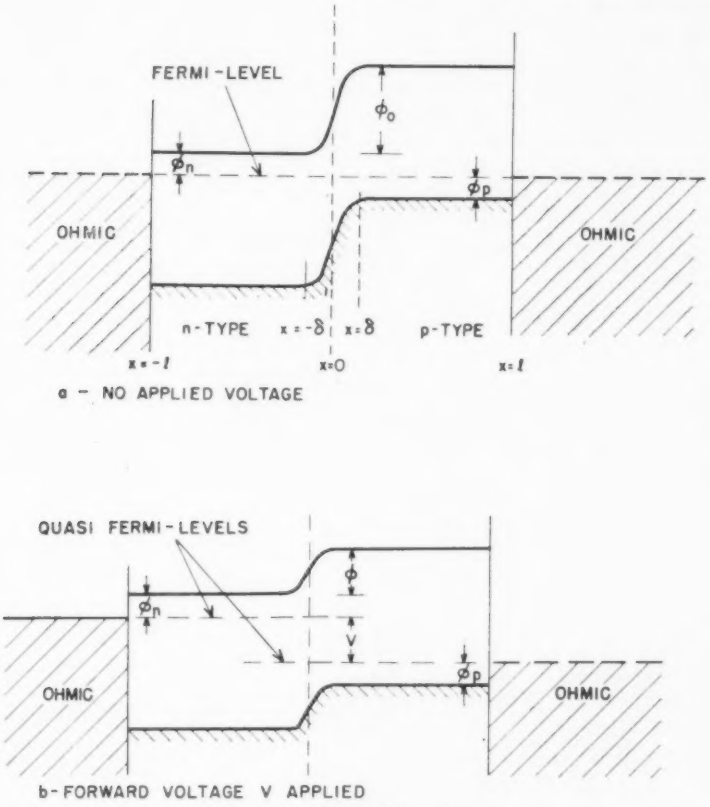


FIG. 1. Isothermal junction.

B. *Junction with thermal gradients*, $\tau = \infty$

We will now discuss the characteristics of the above junction, where the temperatures of the ohmic contacts on the n-type and p-type ends are held at temperatures T_n and T_p respectively, and the junction itself is at T_J. No assumption need be made about the temperature distribution in the regions between $x = 0$ and $x = \pm l$.

By definition, electrons and holes are in equilibrium with each other at the two ohmic contacts, so the minority-carrier concentration at each contact is determined by the corresponding temperature. We will let n_p and p_n be the minority-carrier concentrations at the ohmic contacts at the p-type and n-type ends respectively. Elsewhere in the crystal, the two kinds of carrier will not be in equilibrium with each other, but they will have thermal-energy distributions which correspond closely to the local lattice temperature. Thus, the evaporation of carriers over the potential hill at $x = 0$ will be governed by a Boltzmann factor involving the temperature at the junction.

Equations analogous to (1) and (2) for the electron current densities are

$$J_{e.\,\text{right}} = \frac{qDn_n}{l} \exp(-q\phi/kT_J) \tag{5a}$$

$$J_{e.\,\text{left}} = \frac{qDn_p}{l} \tag{5b}$$

Similar equations describe the flow of holes.

These equations omit terms which describe that component of current flow which is due directly to the thermal gradient. As shown by JOFFE[3], these terms may be either positive or negative depending upon the dominant scattering mechanism, and are small enough to be neglected except in degenerate or nearly degenerate semiconductors.

We will make the further simplifying approximation that the diffusion coefficient is constant throughout the semiconductor and is given by $D = kT_J\mu/q$ where μ is independent of temperature. ϕ is the potential hill at $x = 0$ as determined by the various temperatures and the voltage V applied between the ohmic contacts. As may be seen by Fig. 2,

$$\phi = E_g - \phi_n - \phi_p - V \tag{6}$$

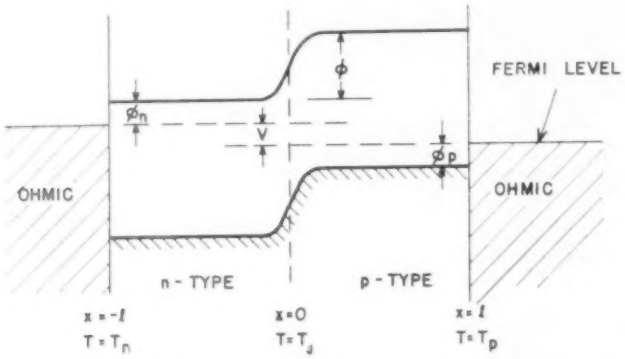


FIG. 2. Junction with thermal gradients.

Recognizing that

$$\frac{E_g}{2} - \phi_n = \frac{kT_n}{q} \ln\left(\frac{n_n}{n_{in}}\right) \tag{7}$$

etc., we get

$$\exp(-q\phi/kT_J) = \left(\frac{n_{in}}{n_n}\right)^{T_n/T_J} \times \times \left(\frac{n_{ip}}{p_p}\right)^{T_p/T_J} \exp(qV/kT_J) \tag{8}$$

The quantities n_{in} and n_{ip} are the intrinsic carrier concentrations corresponding to the temperatures T_n and T_p.

The total current density in terms of the applied voltage is obtained by adding equations (5), and using (8).

$$J = \frac{qD}{l}\left\{\left(\frac{n_{in}n_{ip}}{p_p} + \frac{n_{in}n_{ip}}{n_n}\right)\left(\frac{n_{in}}{n_n}\right)^{(T_n-T_J)/T_J} \times \right.$$
$$\left. \times \left(\frac{n_{ip}}{p_p}\right)^{(T_p-T_J)/T_J} \exp(qV/kT_J) - \left(\frac{n_{ip}^2}{p_p} + \frac{n_{in}^2}{n_n}\right)\right\} \quad (9)$$

This equation is of the form

$$J = \frac{qD}{l}[a \exp(qV/kT_J) - b] \quad (10)$$

When $T_n = T_p = T_J$, we get $a = b$, in agreement with equation (3).

the p-type end of the junction will always develop a positive open-circuit voltage, regardless of the direction of the thermal gradient.

The performance may be characterized by an open-circuit voltage

$$V_{oc} = \frac{kT_J}{q} \ln \frac{b}{a} \quad (11)$$

and a short-circuit density

$$J_{sc} = \frac{qD}{l}(b-a) \quad (12)$$

The maximum power per unit area available to a resistive load may be written

$$P = \alpha V_{oc} J_{sc} \quad (13)$$

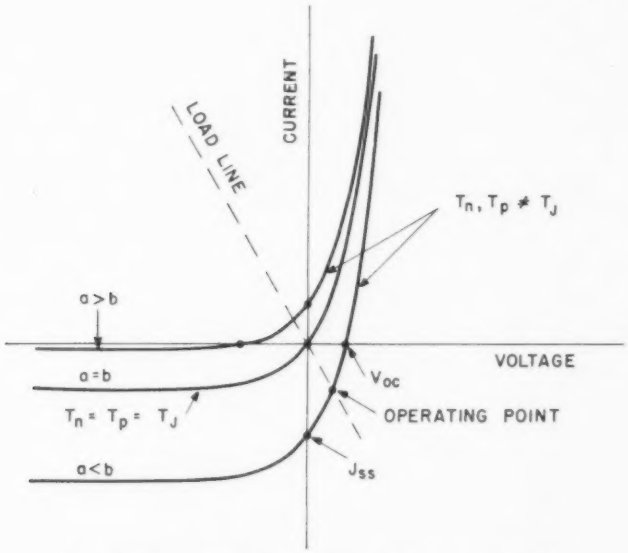


FIG. 3. Current–voltage characteristic of p–n-junction thermoelectric generator.

C. *Thermoelectric-power generation*

The electrical characteristics described by equation (9) are similar to those of the isothermal junction, except that they do not, in general, pass through the point $V = 0$, $J = 0$. In Fig. 3, the upper curve corresponds to $a > b$, while the lower curve corresponds to $a < b$. Inspection shows that cases where $a > b$ are rather uncommon and always give rise to poor thermoelectric efficiency. We will henceforth consider only those situations where $a < b$. Under these circumstances, α depends upon the curvature of the characteristic, but will generally be slightly larger than $\frac{1}{4}$. If the thermocouple characteristic has a sharp curvature in the lower right quadrant of Fig. 3, α may in principle approach unity. Values of α appreciably greater than $\frac{1}{4}$ will be difficult to realize in practice, however, because of the large temperature differences required.

An upper limit to the power output of the thermocouple may be obtained by evaluating the conductance at $V = V_{oc}$. This conductance is

given by

$$G_{oc} = \frac{q\mu b}{l} \tag{14}$$

If the conductance were independent of voltage, the available power would be given by $\frac{1}{4}V_{oc}G_{oc}$. It is easily shown that the maximum power available from a device having the characteristics given by equation (10) is bracketed by the limits

$$\tfrac{1}{4}V_{oc}J_{sc} < P < \tfrac{1}{4}V_{oc}G_{oc}$$

or

$$\frac{q\mu(b-a)}{4l}\left(\frac{kT_J}{q}\right)^2 \ln\frac{b}{a} < P < \frac{q\mu b}{4l}\left(\frac{kT_J}{q}\right)^2\left(\ln\frac{b}{a}\right)^2 \tag{15}$$

For b/a close to unity, both limits converge to the same value. When $b/a \gg 1$, they differ by approximately a factor $\ln b/a$.

The thermoelectric properties of the long-lifetime junction may be illustrated by considering the following cases:

I. $\quad n_n = p_p, \qquad T_n = T_p > T_J$

II. $\quad n_n = p_p, \qquad T_n - T_J = T_J - T_p$

Case I involves the temperature distribution that is normally used with short-lifetime thermocouples, because of the higher thermoelectric efficiency which results. This same distribution could, in principle, be used with long-lifetime junctions. In practice, however, the requirement that the length of the semiconductor be small compared with a diffusion length will probably necessitate operation with a junction temperature approximately half way between the temperatures of the end contacts, as in case II. In each case let $\Delta T = T_n - T_J$. Upper limits to the available power may be calculated from equations (9) and (15). For case I we find

$$P_{\mathrm{I}} < \frac{eq\mu n_{in}}{l}\left(\frac{k\Delta T}{q}\right)^2 \frac{n_{in}}{n_n}\left(\ln\frac{n_n}{n_{in}}\right)^2 \tag{16}$$

This expression has a maximum for $n_n = e^2 n_{in}$ given by

$$P_{\mathrm{I.max}} < \frac{8}{e^2}\frac{q\mu n_{in}}{l}\left(\frac{k\Delta T}{q}\right)^2 \tag{17}$$

For case II, if we take $\Delta T > 0$ and neglect $\ln 2$ compared with $\ln(n_n/n_{in})$, we obtain

$$P_{\mathrm{II}} < \frac{q\mu n_{in}^2}{4ln_n}\left(\frac{E_g\Delta T}{T_n}\right)^2 \tag{18}$$

The greatest power output is obtained when the hot end is nearly intrinsic, so that

$$P_{\mathrm{II.max}} < \frac{1}{4}\frac{q\mu n_{in}}{l}\left(\frac{E_g\Delta T}{T_n}\right)^2 \tag{19}$$

Comparing equations (17) and (19), we see that $P_{\mathrm{II}} > P_{\mathrm{I}}$ whenever $E_g > 2{\cdot}08(kT_n/q)$. Since this will almost certainly be the case, we may conclude that the greatest thermoelectric efficiency will result when the junction temperature is intermediate between that of the two ends. Fortunately, this is the distribution that is most conveniently obtained in practice.

A similar analysis, applied to asymmetrical junctions ($n_n \neq p_p$), gives results that do not differ significantly from equations (17) and (19).

The physical processes which give rise to the thermoelectric voltage of a p–n junction are easily recognized. In Fig. 2, for example, assume that the left ohmic electrode is hotter than the junction, while the other electrode is colder. The principal contribution to the voltage arises from minority carriers (holes) which are generated at the hotter electrode and which diffuse to the junction and are captured by "falling" up the potential hill. Minority carriers from the opposite end behave similarly, and add to this voltage but only by a negligibly small amount because they are present in much smaller concentrations. The contributions due to majority carriers which evaporate in each direction over the potential hill are of intermediate magnitude. Both of these components tend to decrease the thermoelectric voltage.

Clearly, the basic process involved is the generation of one type of carrier in large concentrations at the hot electrode in close proximity to a suitable sink (the potential hill of the junction) to which they can diffuse. The open-circuit potential which is developed is analogous to the floating potential of the collector junction of a transistor, except that the excess minority-carrier concentration is caused by an increase in the temperature of the "emitter" junction instead of by the application of a voltage. This viewpoint suggests the use of a p–n–p or n–p–n structure as being inherently simpler and

possibly more efficient than the ohmic–*n*–*p*–ohmic configuration which was analyzed in subsection B. Since the thermoelectric efficiency was seen to increase as the intrinsic temperature was approached, it is natural to inquire into the efficiency of a structure wherein the end regions are very strongly doped and are separated by a region of intrinsic semiconductor of thickness $2l$. An analysis similar to the preceding ones shows that for this structure

$$P < \frac{1}{4\sqrt{(2)}} \frac{q\mu n_i}{l} \left(\frac{E_g \Delta T}{T}\right)^2 \tag{20}$$

A temperature difference, $2\Delta T$ is maintained between the two strongly doped regions and it is assumed that no recombination takes place in the intrinsic region. In order to obtain a solution, it was necessary to assume a finite lifetime in the end regions, but the actual value assumed does not affect the result. T is the temperature of the hot junction, and n_i is the corresponding intrinsic concentration. It is assumed that ΔT is large enough that the intrinsic concentration at the cold end is negligible compared with that at the hot end. Comparison with equation (19) shows that the output is smaller by a factor of $\sqrt{(2)}$ than that of the *p*–*n* junction, operated as in case II.

For purposes of comparison, we note that the open-circuit voltage of a conventional thermocouple made from an extrinsic semiconductor is given by[3]

$$V_{oc} = \frac{k\Delta T}{q}\left[m + \ln\left(\frac{N_c}{n}\right)\right], \quad (2 < m < 4) \tag{21}$$

and the conductance by $G = nq\mu/l$. The power available to a matched load, $P_{av} = \frac{1}{4}GV_{oc}^2$ has a maximum for $n = N_c \exp(m-2)$, (assuming constant mobility) and the corresponding output power is

$$P_{\max} = e^{m-2} \frac{q\mu N_c}{l} \left(\frac{k\Delta T}{q}\right)^2 \tag{22}$$

For future reference we may note that the "figure of merit" under these circumstances is given by

$$z = 4\frac{k^2\mu}{q\kappa} N_c \exp(m-2) \tag{23}$$

In order to compare the performances of the thermoelements which have been considered, we will express the results in terms of the temperature difference δT between the hot and cold ends of the device, and divide by the thermal energy transported by lattice conduction. In this way we are able to take into account explicitly the fact that heat is transported by both halves of the junction in parallel in case I, and that in case II, $\delta T = 2\Delta T$. These thermoelectric efficiences do not take into account thermal transport by free carriers or the effects of joule heating. They should, however, serve as a fair basis for comparing the relative performance capabilities of the various kinds of devices. It is also helpful to write $n_{in} = N_c \times \times \exp(-qE_g/2kT)$. ($T$ refers to the hot-end temperature.) The results are as follows:

p–n junction with ends at same temperature (*case* I):

$$\eta_{\mathrm{I}} = \frac{4}{e^2} \frac{k^2\mu N_c}{q\kappa} [\exp(-qE_g/2kT)]\delta T \tag{24}$$

p–n junction with ends at different temperatures (*case* II):

$$\eta_{\mathrm{II}} = \frac{1}{8} \frac{k^2\mu N_c}{q\kappa} \left(\frac{qE_g}{kT}\right)^2 [\exp(-qE_g/2kT)]\delta T \tag{25}$$

Short-lifetime thermocouple:

$$\eta_{th} = e^{m-2} \frac{k^2\mu N_c}{q\kappa} \delta T \tag{26}$$

κ is the thermal conductivity.

The most important difference between these expressions is the factor $\exp(-qE_g/2kT)$. Because of this fact, it may be concluded that long-lifetime thermoelectric generators must have poor efficiencies unless constructed from very small-band-gap semiconductors. Under these circumstances, many of the approximations made in deriving the formulae are invalid, but qualitative considerations would seem to rule out the possibility of achieving good performance from small-band-gap semiconductors as well.

A physical argument can also be used to show rather generally that a long carrier lieftime will not contribute substantially to the performance of thermoelectric devices. Changes in lifetime can only be expected to have an appreciable effect on the concentrations of minority carriers. Therefore large minority-carrier concentrations are required

if this change in lifetime is to produce an appreciable change in the efficiency of the thermoelectric device. This, in turn, implies the use of smaller majority-carrier concentrations than are normally used in semiconductor thermoelements and hence a rather high electrical impedance and a correspondingly small power output. In other words, the high electrical impedance which is inherent in a p–n junction imposes a severe limitation upon the current density that can be achieved in a thermoelectric generator containing a junction.

D. *Thermoelectric cooling*

When a forward current is passed through a p–n junction, electrons and holes evaporate over the potential hill at the junction, and eventually recombine, releasing energy at some distance from the junction. In the device described by equation (9), the energy which is absorbed at the junction is liberated by recombination at the ohmic electrodes, and it might appear that such a junction could provide an efficient means of refrigeration. We will now calculate the temperature difference $\Delta T = T_n - T_J$ that can be obtained with a long-lifetime junction, operated as a thermoelectric cooler, with $T_n = T_p$ and $n_n = p_p$, using equation (9).

The heat extracted at the junction is given by

$$Q_{\text{ext}} = J\phi = \frac{2kT_nJ}{q}\ln\left(\frac{n_n}{n_{in}}\right) - JV \qquad (27)$$

Since heat is supplied by thermal conduction at a rate $2\kappa\Delta T/l$, the net rate at which heat is removed is given by

$$Q = 2\frac{kT_nJ}{q}\ln\left(\frac{n_n}{n_{in}}\right) - JV - \frac{2\kappa\Delta T}{l} \qquad (28)$$

Equations (9) and (28) may be used to obtain a single relation between the three variables, Q, J and ΔT.

For sufficiently small values of ΔT and J, a first-order approximation shows that the coefficient of performance approaches the thermodynamic limit of $T_J/\Delta T$. At higher current densities, we may neglect b in comparison with $a\exp(qV/kT_J)$ in equation (10), and obtain a result which is not limited to small values of ΔT.

$$Q = \frac{JkT_J}{q}\ln\left(\frac{2kT_J\mu n_n}{Jl}\right) - \frac{2\kappa\Delta T}{l} \qquad (29)$$

The rate at which heat is removed from the junction is greatest at a current density of $J = 2kT_J\mu n_n/el$, giving

$$Q_{\max} = \frac{2k^2T_J^2\mu n_n}{eql} - \frac{2\kappa\Delta T}{l} \qquad (30)$$

The maximum temperature difference that can be obtained occurs when $Q_{\max} = 0$.

$$\Delta T_{\max} = \frac{k^2T_J^2\mu n_n}{eq\kappa} \qquad (31)$$

This may be compared with the maximum cooling obtainable with a conventional thermocouple, using equation (23).[3]

$$\Delta T_{\max} = \frac{zT_J^2}{2} = 2e^{m-2}\frac{k^2T_J^2\mu N_c}{q\kappa} \qquad (32)$$

Here, again, the analysis indicates that cooling may be accomplished more efficiently with conventional thermocouples than with long-lifetime p–n junctions.

4. CONCLUSIONS

In the foregoing sections, analyses were made of the thermoelectric properties of several semiconductor structures having characteristics dependent upon the carrier lifetime. In each case, the performance was distinctly inferior to that of a conventional short-lifetime thermocouple. Considerations based upon the physical processes involved make it appear very unlikely that an increase in carrier lifetime can be used to provide an improvement in either thermoelectric power generation or in cooling.

Acknowledgements—The author is indebted to L. Apker and H. Ehrenreich for a number of criticisms and helpful suggestions.

REFERENCES

1. J. Tauc, *Czech. J. Phys.* **6,** 108 (1956).
2. C. Herring, *Phys. Rev.* **96,** 1163 (1954).
3. A. F. Joffe, *Semiconductor Thermoelements and Thermoelectric Cooling*. Infosearch Ltd., London (1957).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 123–132. Printed in Great Britain

DETAILED ANALYSIS OF THIN PHOSPHORUS-DIFFUSED LAYERS IN *p*-TYPE SILICON

EILEEN TANNENBAUM

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey

(*Received* 20 *October* 1960)

Abstract—Detailed concentration profiles of phosphorus-diffused layers in *p*-type silicon have been obtained using a new, accurate technique. The phosphorus distributions were found to be significantly different from those predicted by simple theory in that: (1) under conditions of constant surface concentration, they do not obey the predicted error-function complement, (2) the diffusion coefficient is a constant up to a concentration of about 10^{20} cm^{-3}, but is a strong function of concentration above that value, and the variation is larger than can be accounted for by field-accelerated diffusion and (3) at high concentrations, there is a difference between the phosphorus concentration determined from resistivity measurements and that determined from radioactive tracer measurements.

Electron mobility values in the concentration range above 10^{20} cm^{-3} were obtained from the data. The experimental results suggest that there is a limit to the concentration of electrically active phosphorus that can be obtained in silicon, and that this limit has been observed.

Résumé—Des profils détaillés de la concentration de phosphore en de couches diffusées dans le silicium de type *p* ont été obtenus en employant une nouvelle méthode précise. Les distributions de phosphore obtenues ont été très différentes de celles qu'on aurait prédit théoriquement en ce que: (1) sous des conditions de concentration constante en surface, elles ne varient pas comme la fonction d'erreur, (2) le coefficient de diffusion est une constante jusqu'à une concentration d'environ 10^{20} cm^{-3}, mais est en fonction de la concentration au-dessus de cette valeur, et la variation est plus grande que celle que pourrait occasionner la diffusion accélerée par le champ et (3) aux hautes concentrations, il y a une différence entre la concentration de phosphore déterminée par des mesures de resistivité et celle obtenue par des mesures de tracement radioactif.

Des données, on a obtenu des valeurs de mobilité d'électrons dans la gamme de concentration au-dessus de 10^{20} cm^{-3}. Les résultats expérimentaux démontrent qu'il y a une limite à la concentration du phosphore électriquement actif qui peut-être obtenu dans le silicium et que cette limite a été observée.

Zusammenfassung—Durch ein neues und exaktes Verfahren wurden detaillierte Dichteprofile von Schichten diffundierten Phosphors in *p*-Typ Silizium untersucht. Es ergab sich, dass die Verteilung des Phosphors sich wesentlich von der Erwartung der einfachen Theorie unterschied: (1) Bei konstanter Oberflächenkonzentration genügten sie dem erwarteten Ausgleich der Fehlerfunktion nicht. (2) Bis zu Konzentrationen von etwa 10^{20} cm^{-3} ist der Diffusionskoeffizient konstant; oberhalb dieses Wertes hängt er stark von der Konzentration ab, und die Variation ist stärker als auf Grund einer durch ein Feld beschleunigten Diffusion zu erwarten wäre. (3) Bei hohen Konzentrationen ergibt sich ein Unterschied zwischen der durch Messung des spezifischen Widerstands bestimmten Phosphorkonzentration und der durch radioaktive Indikatoren bestimmten.

Aus den Ergebnissen wurden Werte der Elektronenbeweglichkeit im Konzentrationsbereich über 10^{20} cm^{-3} bestimmt. Die experimentellen Ergebnisse deuten an, dass die Konzentration von elektrisch aktivem Phosphor, die in Silizium erzielt werden kann, eine Grenze hat, und dass diese Grenze beobachtet wurde.

1. INTRODUCTION

For some time it has been apparent that sizable discrepancies exist between experiment and theory in the diffusion of electrically active impurities in silicon. The discrepancies often appeared in the form of inaccuracies of the predicted junction

depth, and apparent variations in diffusion coefficient with background doping particularly for shallow diffusions. It therefore seemed worthwhile to look at the distribution of impurities in diffused layers. The system chosen for study was phosphorus diffused into *p*-type silicon. Although the method of analysis is not new in principle, accurate, detailed distributions were obtained by the use of a new, experimental technique. The technique differs from the conventional one only in the method of layer removal. Instead of lapping, layers are removed by anodic oxidation. This method not only permits the removal of very small increments but ensures the removal of layers parallel to the surface. The body of the paper describes the results of this study.

2. EXPERIMENTAL DETAILS

A. *Diffusion and sample preparation*

All data were obtained on two boron-doped pulled crystals with a resistivity of about 2 Ω-cm. The prediffusions were carried out in an open-tube diffusion furnace.[1] The term "prediffusion" refers to heating the sample in the presence of the diffusant, i.e. a condition of constant surface concentration which ideally yields an error function complement distribution of the diffusant. The term "diffusion" refers to additional heating in the absence of the diffusant, a condition which ideally yields a Gaussian distribution. Some samples were analyzed after the prediffusion, and others (after removing the glassy P_2O_5-layer with HF, and cleaning) were further diffused in a furnace which had never been exposed to P_2O_5. The diffusion data are summarized in Table 1.

After diffusion, an island ($\sim 3\ \text{mm} \times 8\ \text{mm}$) was masked on the diffused slice and the exposed diffused area was etched away.

Table 1. Phosphorus diffusion in silicon

I. N_s = constant*

T(°C)	t(hr)	No. samples
800	1·0	1
900	0·5	4
1000	0·5	2
1050	0·5	3

II. $\int_0^\infty N_x dx$ = constant*

Prediffusion		Diffusion		No. samples
T°(C)	t(hr)	T(°C)	t(hr)	
900	0·5	906	6	1
		906	15	1
		906	18	1
1000	0·5	1015	2	1
		1015	3	1
		1015	6	1
		1015	8	2
		1015	15	1
		1015	16	1
		1015	24	1
1000	0·5	1114	2	1
		1114	3	1
		1114	4	1
1050	0·5	1015	4	2
		1015	8	2

* N_s is the surface concentration per cm³, and N_x is the concentration per cm³ in a plane parallel to the surface at a depth x from the surface.

B. *Resistivity and depth measurements*

Thin layers of silicon were removed parallel to the junction by anodic oxidation. The slice was anodized in a 0·1 M boric acid, sodium tetraborate solution. The thickness of the oxide grown in this way was determined from calibration charts of color versus thickness. The color of air films between reflecting plates has been tabulated as a function of thickness.[2] It can be shown that these tables are applicable to SiO_2 films by correcting for the refractive index of bulk silica. The thickness of the Si oxidized is a fraction of the SiO_2 thickness calculated from the molecular weights and densities of Si and SiO_2. In a single anodization, a minimum of ~ 300 Å, and a maximum of ~ 1200 Å Si can be removed in the calibrated range. Since the sensitivity of the color change with thickness varies over the spectrum, the most sensitive range was chosen, and, for the most part, the silicon was removed in increments of about 400 Å ± 50 Å. The depth at any point is then the sum of these increments. There can be no change in the phosphorus

distribution during anodization since it is a room-temperature process, and the phosphorus in the anodized layer is removed by dissolving the oxide in HF. The samples were rinsed in hot deionized water prior to measurement of the resistivity.

The resistivity measurements were made on a four-point probe a.c. semiconductor resistivity bridge developed by LOGAN[3]. The apparatus is capable of a precision of $\pm 0{\cdot}5$ per cent and measurements were reproducible to within that precision. This holds until the anodization has proceeded to within about 2000 Å of the junction. At this point, the sheet resistance is generally in the range of $\sim$ 1000 Ω/square, and the measurements have a tendency to drift somewhat with a reproducibility of approximately 10 per cent. Since it is only the last three or four measurements on a given sample which are subject to drift, an accuracy of 1 per cent in the resistivity measurements can be safely claimed. The depth measurement per anodization step is in error by no more than ± 50 Å. To check the accuracy of this method of depth measurement, a sum of forty-two anodizations was compared with an angle-lap measurement of the junction depth. The two agreed within 0·01 mil, which is the uncertainty in the angle-lap measurement. All the data form smooth curves within the given error limits. Considering the over-all error, including uncertainties in calibration constants as well as uncertainties in the measurement of slopes, the experimental data are believed to be accurate to within 20 per cent.

C. *Radioactive tracer data*

Two samples, prediffused at 1050°C for 30 min, were neutron bombarded for a week to produce radioactive phosphorus. After bombardment, the sheet resistance was found to have increased by about a factor of 2, presumably due to bombardment damage. A brief heat treatment (1050°C, 3 min), followed by quenching, lowered the resistivity back to its former value. The half-life checked with that of phosphorus-32. A blank irradiated sample indicated the presence of surface gold,[4] by checking half-life. Although the gold was probably removed by the anodization, one of the samples was allowed to decay until the estimated gold had decayed to $< 0{\cdot}1$ per cent of the phosphorus count. The analysis was carried out as described above so that, at each depth, a parallel measurement of sheet resistance and total counts of the diffused layer were obtained. The data were corrected for background and decay. The counts were taken for a length of time such that the counting error remained less than 1 per cent throughout.

3. ANALYSIS OF DATA

A. *Calculation of concentration profiles*

The experimental data are plotted in the form of ρ_s^{-1} as a function of x, where ρ_s is the sheet resistance and x is the depth from the original surface. The slope $[d(\rho_s^{-1})/dx]$ at any point on the curve is a measure of the conductance in a plane parallel to the junction passing through the point x. The conductance is related to the concentration by the expression

$$\frac{d(\rho_s^{-1})}{dx} = \sigma_x = N_x \mu_x' q \qquad (1)$$

where N_x is the concentration of carriers (cm^{-3}), q is the electron charge and μ_x' is the effective mobility for electrons.

The effective mobility values used were a compilation by IRVIN[5], of the best values obtained from published and some unpublished results. The mobility data are believed to be accurate to within 20 per cent for concentrations below about 5×10^{19} cm^{-3}. Above this concentration, very few mobility data are available, and μ' is assumed to be a constant, $\mu = 75$ cm^2/V-sec.

B. *Calibration of tracer data*

The tracer data were treated in the same way as resistivity data in that the total counts per minute were plotted as function of x. The slope at any point gives an arbitrary number proportional to the concentration at that point. The ratio of this number to the concentration determined from resistivity measurements was determined at each value of x. The ratio was found to be a constant to within ± 10 per cent at all concentrations below 5×10^{19} cm^{-3} for both samples. The tracer data were thus calibrated using the constant ratio.

C. *Calculation of diffusion coefficients*

The values of the diffusion coefficient shown in Fig. 9 were obtained by a method outlined by CRANK[6]. The method involves the application of Boltzmann's transformation ($\eta = x/2t^{1/2}$) to the

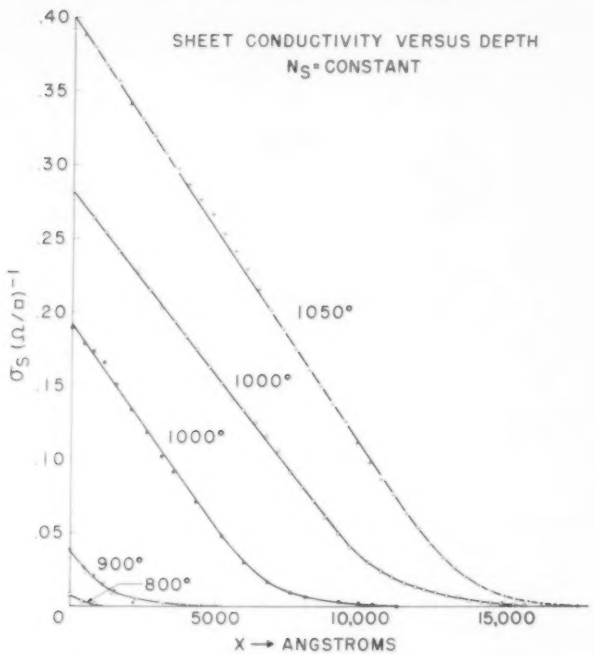


FIG. 1. Sheet conductance as a function of depth for a series of samples diffused under conditions of constant surface concentration. The 900°C curve contains data from three different samples; the other curves each represent a single sample.

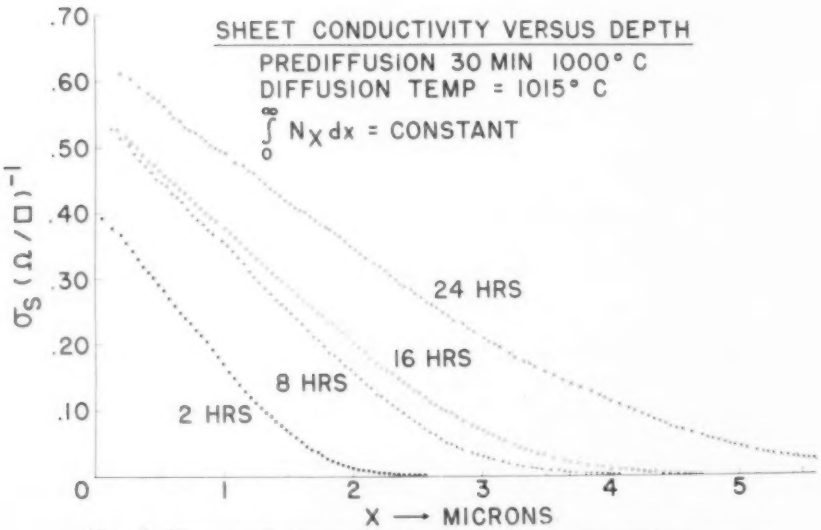


FIG. 2. Sheet conductance as a function of depth for a series of samples diffused for the indicated times at $T = 1015°C$, in flowing oxygen. All samples initially contained a layer prediffused for 30 min at 1000°C.

diffusion equation. The equation then is reduced to an ordinary differential equation which can be integrated with respect to η. Reintroducing x and t gives

$$D_{(N=N_x)} = \frac{1}{2t}\left(\frac{dx}{dN}\right)_{N_x} \int_0^{N_x} x\, dN \qquad (2)$$

This equation is valid when the diffusion coefficient is a function only of concentration and when the initial and boundary conditions can be expressed in terms of η alone, i.e. x and t are not involved separately.

The integrations were done graphically, and the slopes measured directly from the concentration–distance curves, obtained from the tracer measurements.

4. RESULTS AND INTERPRETATION

A. *Discrepancies between experiment and simple theory*

Figs. 1 and 2 give a representative sampling of the experimental data. It is clear by inspection of Fig. 1 that all the data above 900°C form straight lines over a considerable portion of each curve. Since the slope is a constant, then the concentration is a constant assuming that the mobility is a constant over the pertinent resistivity range. Using the mobility data cited previously, concentration profiles are obtained as shown in Figs. 3 and 4 on logarithmic and linear scales, respectively. The commonly assumed error-function complement was calculated for comparison using MACKINTOSH's[8] value of the diffusion coefficient and the experimentally determined junction depth. It can be seen in Fig. 4 that although the

$$\int_0^\infty N_x\, dx$$

for the experimental curve and the erfc agree within reproducibility limits, the surface concentration of *electrically active* impurities, calculated assuming an erfc distribution, is far higher than that found experimentally. In addition, the observed concentration remains constant for an appreciable depth rather than decreasing exponentially from the surface.

Fig. 5 shows concentration profiles, calculated

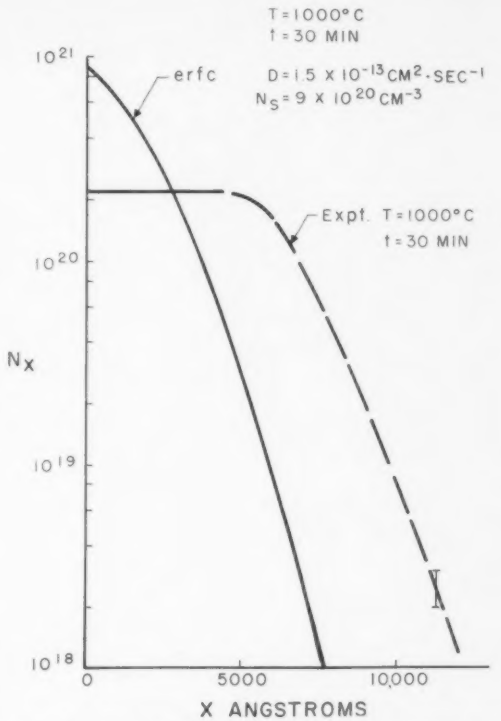


FIG. 3. Logarithm of the calculated concentration as a function of depth. The vertical line indicates the maximum deviation of the calculated points from the dashed curve. The complementary error function, predicted by simple theory, is shown for comparison.

as outlined above, from the data shown in Fig. 2. The vertical line on the upper portion of the curve marked "2 hr" indicates the maximum deviation of any calculated point from the smooth curve, and the small vertical line indicates the typical variation. The distributions labeled 2 hr, 8 hr, etc., were obtained by diffusion from layers prediffused at 1000°C for 30 min with a distribution indicated in Fig. 5. It may be noticed that there appears to be more phosphorus in the diffused samples than was present in the original prediffused layer. Integrating these curves and taking the ratio of the number of carriers in the diffused layer to the number of carriers in the prediffused layer gives the results shown in Fig. 6. The ratio increases rapidly, approaching about twice the original total apparent phosphorus. The only possible interpretation of this result is that

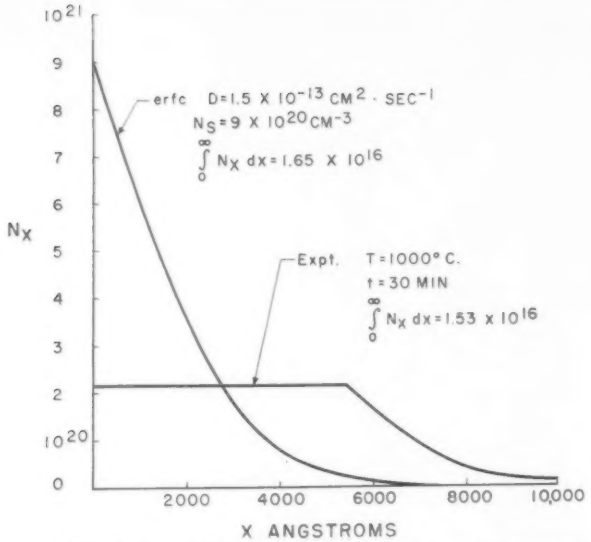


FIG. 4. Linear plot of the calculated concentration as a function of depth. The data are the same as in Fig. 3. Note that, although the distributions are widely different, the integrals of the theoretical and experimental curves are very nearly the same.

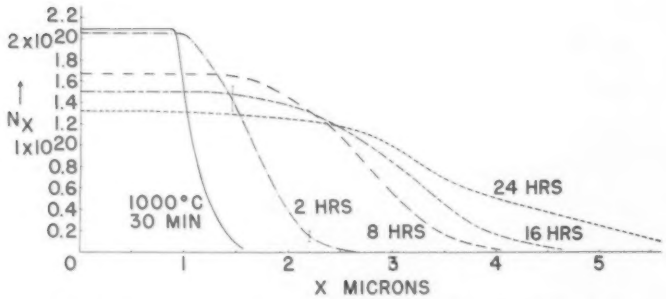


FIG. 5. Linear plot of the distributions calculated from the experimental data shown in Fig. 2. The solid curve represents the distribution of phosphorus prior to diffusion. Note the increase in the integral of each curve as a function of time.

the resistivity measurements do not indicate the *total phosphorus* present either because of "precipitation"* of phosphorus (i.e. "electrically inactive" phosphorus) or because of an error in the mobility at high concentrations, or both.

* Precipitated phosphorus for the purposes of this discussion is defined as phosphorus so situated in the lattice that it cannot contribute to the conductivity. This is not to be confused with un-ionized phosphorus.

The tracer experiments were undertaken to obtain a direct comparison of the total phosphorus present and the electrically active carriers obtained from resistivity measurements. The comparison is shown in Fig. 7 for one of the samples. The ratio of the integrals of the two curves in Fig. 7 is roughly 2. Referring back to Fig. 6, the ratio of the phosphorus in the diffused layer to phosphorus in the prediffused layer is also 2 for long diffusion

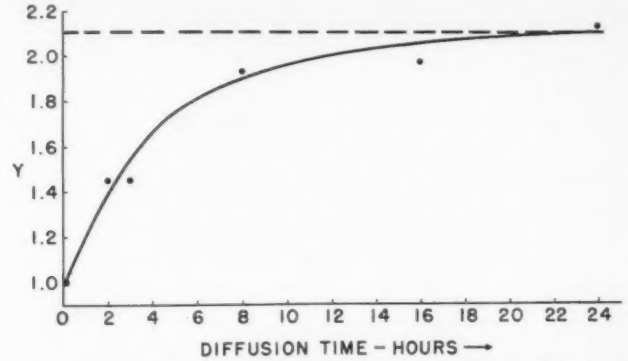


FIG. 6. Y is the ratio of the integral of the distribution obtained after diffusion for a time t to that of the original prediffused layer. Note that the total apparent phosphorus concentration increases as a function of time and approaches a constant.

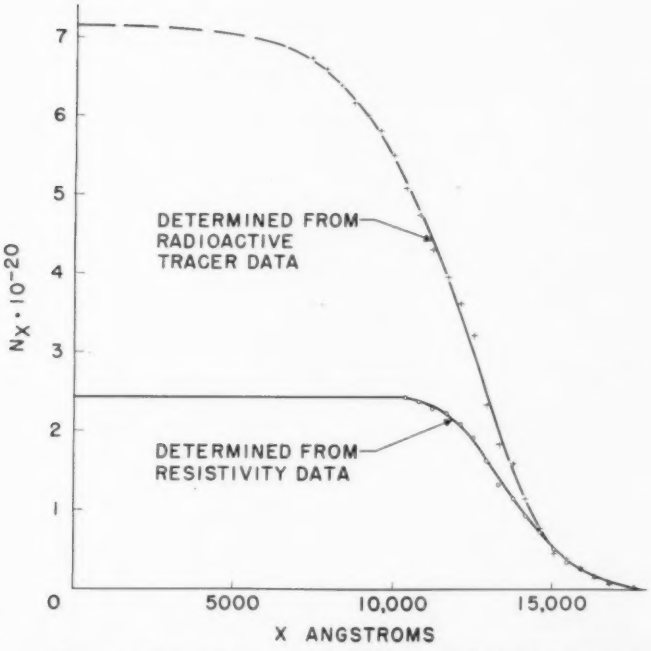


FIG. 7. Comparison of the distribution obtained by tracer measurements (total phosphorus) with that obtained from resistivity measurements (electrically active phosphorus). Both sets of measurements were obtained on the same sample, prediffused for 30 min at 1050°C.

times. Therefore, after long diffusion times from a prediffused layer, essentially all the phosphorus is detected by the resistivity measurements.

If one were arbitrarily to attribute the entire difference between the tracer and the resistivity results to a mobility correction, then one finds that

the mobility continues to decrease with increasing concentration (rather than approach a constant) and in fact becomes inversely proportional to the concentration at the point where the concentration (determined from resistivity) levels off. See Fig. 7. The effect of the above assumption on the concentration versus resistivity plot is shown in Fig. 8. The crosses in Fig. 8 are the compiled data of

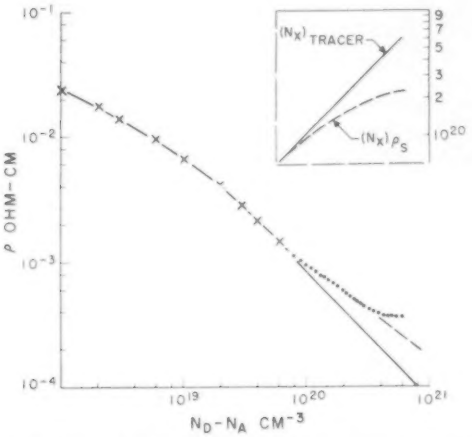


FIG. 8. Resistivity as a function of net donor concentration. The crosses are the compiled data of IRVIN[5]. At concentrations above 6×10^{19} cm^{-3} it is generally assumed that $\mu = 75$ cm^2/V-sec, giving a straight line extrapolating down to 8×10^{20} cm^{-3} at $\rho = 10^{-4}$ Ω-cm, on this plot. The dots are the result of calculating values of the mobility which force the distribution obtained from resistivity measurements to agree with that obtained from the tracer measurements. The inset shows the divergence of the two concentration determinations at high concentrations. The abscissa of the inset is the same as that of the graph.

IRVIN[5], and, above a concentration of $\sim 6\times10^{19}$, the mobility is assumed to be a constant ($\mu' = 75$ cm^2 V$^{-1}\times$sec^{-1}). This would give a straight line for the plot of (N_D-N_A) versus ρ extrapolating down to $\sim 8\times10^{20}$ cm^{-3} at $\rho = 10^{-4}$ Ω-cm. Recent mobility data[9,10] in the concentration range from 5×10^{19} to $1{\cdot}5\times10^{20}$ cm^{-3} indicate slightly lower values than IRVIN's assumption, but within his quoted accuracy. These new data do not significantly affect the results obtained in this investigation. The circles indicate the result of changing the mobility in such a way as to make $(N_x)\rho_s$ agree with (N_x) tracer.

One interpretation of the graph is that μ is constant, $\mu = (75$ cm^2/V-sec), and that the entire deviation of the points from the extrapolated line is due to precipitation. Another possibility is that the mobility continues to decrease, falling to ~ 50 cm^2/V-sec at a concentration of 3×10^{20} cm^{-3} and that the leveling of the curve beyond this point indicates the onset of precipitation. It is impossible to choose between these two interpretations on the basis of the experiments reported here. However, the conclusion that some precipitation is taking place seems unavoidable.

It is generally assumed that, in cooling rapidly from diffusion temperature to room temperature, a supercooled solution is obtained with substantially all the phosphorus frozen into substitutional positions. However, at high concentrations, precipitation can occur upon cooling from diffusion temperatures depending on the rate of cooling. To check this point, a sample diffused for 30 min at 1050°C was subjected to a series of heating and cooling cycles to observe the effect on the resistance*. The results are presented in Table 2 in the order in which the series was carried out.

Table 2

Treatment	ρ_s(Ω/square)
1. Pulled from diffusion furnace, allowed to cool in room air on quartz platform.	2·73
2. Heated to 1050°C for 3 min, pulled from furnace and dropped into oil bath.	2·30
3. Heated to 1050°C for 3 min, cooled over 3-min period by pulling slowly from furnace.	2·50

These results are in qualitative agreement with the precipitation model. However, this result is not conclusive since the resistance changes observed may also be ascribed to other causes such as strains and dislocations.

One can hypothesize, on the basis of these results, that there is a limit to the amount of electrically active phosphorus that can be obtained in

* I wish to thank F. M. SMITS for suggesting this experiment.

silicon. It is noteworthy that, from 900 to 1050°C, the surface concentration of electrically active carriers (determined from resistivity measurements and assuming $\mu = 75$ cm^2/V-sec) is independent of temperature and is equal to $2{\cdot}2 \pm 0{\cdot}2 \times 10^{20}$ cm^{-3}.

B. *Determination of diffusion coefficients*

It is apparent from examination of the tracer profile that the distribution is not an erfc. One probable reason for a deviation from the erfc is that the diffusion coefficient is not a constant. Therefore, the diffusion coefficient was calculated as a function of concentration by the method outlined in Section IIIC, using the distributions obtained from the tracer data. The results are shown in Fig. 9. Within experimental error, the diffusion coefficient is a constant at concentrations below

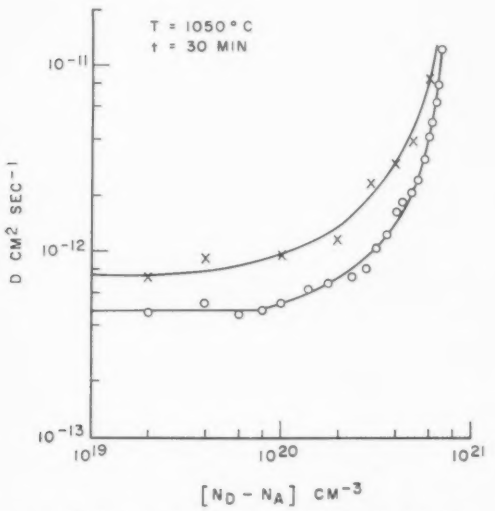


FIG. 9. Diffusion coefficient at 1050°C calculated from equation (2) using the distributions obtained from the radioactive-tracer data. Crosses and circles represent two different samples diffused under the same conditions.

about 10^{20} cm^{-3}. Above that concentration, the diffusion coefficient rises very rapidly. This variation of diffusion coefficient is too large to be attributed to the effect of diffusion potentials which yields only a factor of 2 change in the diffusion coefficient.[7]

The low-concentration limit of the diffusion coefficient was calculated for the distributions at 900 and 1000°C using the same method as above. The results are shown in Table 3.

Table 3

T°C	$Dn_x < 10^{20}$cm^{-3}
$900 \pm 10°$	$1{\cdot}7 \times 10^{-14} \pm 20$ per cent cm^2/sec
$1000 \pm 10°$	$4 \times 10^{-13} \pm 20$ per cent
$1050 \pm 10°$	$6 \times 10^{-13} \pm 20$ per cent

These values agree with those obtained by MACKINTOSH[8] within diffusion control limits.

5. SUMMARY

An analysis of the distribution of thin phosphorus-diffused layers in silicon has led to the following conclusions:

(1) The distribution is not an erfc for $T > 900$°C and $N_s > 10^{20}$ cm^{-3}.
(2) The diffusion coefficient at a constant temperature is constant up to concentrations of about 10^{20} cm^{-3} and increases rapidly at higher concentrations.
(3) Sheet-resistance measurements do not give a true picture of the total phosphorus in the sample at high concentrations, but rather indicate only the concentration of "electrically active" phosphorus.
(4) The mobility is either a constant ($\mu = 75$ cm^2/V-sec) above 8×10^{19} cm^{-3}, with phosphorus precipitating above that concentration, or the mobility continues to decrease to a value of about 50 cm^2/V-sec at a concentration of about 3×10^{20} cm^{-3} with phosphorus precipitating above that concentration.
(5) There appears to be a lower limit to the resistivity obtainable at room temperature for phosphorus-doped silicon and that limit is $\sim 0{\cdot}00035$ Ω-cm.

Acknowledgements—I wish to thank Mrs. D. R. KETCHOW and J. F. ASCHNER for providing some of the diffused samples and especially E. I. POVILONIS for the formidable task of anodization and resistivity measurements.

REFERENCES

1. B. T. HOWARD, *Fall Meeting, Electrochemical Society, Buffalo, October*, 1957.
2. J. ROLLET, *S.B. Akad. Wiss. Wien* **57,** 299 (1878).
3. M. A. LOGAN, *Bell Syst. Tech. J.* In press (1961).
4. G. BEMSKI and J. D. STRUTHERS, *J. Electrochem. Soc.* **105,** 588 (1958).
5. J. C. IRVIN. Private communication (1960).
6. J. CRANK, *The Mathematics of Diffusion* pp . 148, 233 Clarendon Press, Oxford (1956).
7. F. X. HASSION and L. J. RUSSO. Private communication (1956).
8. I. M. MACKINTOSH, *J. Electrochem. Soc.* To be published (1961).
9. L. ESAKI and Y. MIYAHARA, *Solid-State Electron.* **1,** 13 (1960).
10. R. A. LOGAN, J. F. GILBERT and F. A. TRUMBORE. Private communication (1960).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 133–142. Printed in Great Britain

INDIUM ARSENIDE TUNNEL DIODES

HANS P. KLEINKNECHT*

RCA Laboratories, Princeton, New Jersey

(*Received* 10 *August* 1960)

Abstract—The tunnel-current density through a very narrow p–n junction depends exponentially on the reduced effective mass of the charge carriers and the bandgap. Therefore, indium arsenide appears to be a promising material for fast tunnel diodes at room temperature ($m_r \simeq 0{\cdot}05$, $E_g \simeq 0{\cdot}33$ eV), giving higher frequency limits for a given doping level or lower doping (lower capacity) for a given gain-bandwidth product.

Tunnel diodes have been made by alloying In–Cd and In–Zn dots on InAs crystals which had been doped with 2×10^{16} to 6×10^{18} Te atoms/cm^3. The resulting I–V characteristics vary with doping level and temperature from classical p–n curves to Esaki curves of 10:1 peak to valley current ratios. The best ratios at room temperature are about 2:1.

Good units with high current density do not show any structure indicative of phonon participation. However, lower-current units do show two conductance minima in the rising part of the forward characteristic at 4°K. The high-current units also show a different temperature dependence of the peak current and of the peak voltage than those of lower current density. This leads to the tentative conclusion that the characteristics of the low-current units are determined by phonon-assisted tunneling, while in the high-current units the direct tunneling process is dominant.

In the I–V characteristic of units made from low-doped material an interesting cusp structure is found. This resembles the "phonon" structure above, but is dependent on the surface condition of the diode.

A measurement of the rise time of a switching pulse across the diode proved to be limited by the circuit to 10^{-9} sec. This and other indirect measurements allow an estimate of a gain-bandwidth product, for the faster diodes, of 3 kMc/s.

Résumé—La densité du courant tunnel à travers une jonction p–n très étroite est en fonction exponentielle de la masse réduite effective de la charge des porteurs et de la largeur de la bande interdite. Donc, l'arséniure d'indium semble être un matériau très intéressant pour les diodes tunnel rapides à température ambiante ($m_r \simeq 0{,}05$; $E_g \simeq 0{,}33$ eV) donnant de plus hautes limites de fréquence pour un certain niveau de dopement, ou moins de dopement (plus basse capacité) pour un certain produit gain-largeur de bande.

Des diodes tunnel ont été fabriquées par un alliage de points d'In–Cd et d'In–Zn sur des crystaux de InAs dopés d'un nombre d'atomes de Te allant de 2×10^6 à 6×10^8 atomes/cm^3. Les caractéristiques (I–V) qui en résultent varient en fonction du niveau de dopement et de la température, entre les courbes p–n classiques et les courbes d'Esaki de 10:1 (rapport du courant de crête au courant de "vallée"). Les meilleurs rapports à température ambiante sont de 2:1.

Les bons produits à densité de courant élevée ne démontrent aucune structure indiquant une participation de phonons. Pourtant, les produits à courant plus bas se distinguent par deux minima de conductance dans la partie avant de la courbe caractéristique à 4°K. Les produits à densité de courant plus élevée se distinguent aussi par le fait que leur courant et voltage de crête ne sont pas tributaires de la température au même degré que les produits à densité de courant plus basse. Ceci mène à tenter de conclure que les caractéristiques des produits à basse densité de courant sont déterminées par un passage à tunnel aidé de phonons, tandis que pour les produits à densité de courant élevée le procédé de passage à tunnel direct est prépondérant.

Dans les caractéristiques (I–V) des produits fabriqués avec des matériaux légèrement dopés, une intéressante cuspide a été découverte. Celle-ci ressemble à la structure de phonons décrite plus haut, mais dépend de la condition de la surface de la diode.

* Present address: Intermetall G.m.b.H. Hans-Bunte-Strasse 19, Freiburg i.B., Germany.

La mesure du temps de relèvement d'une impulsion commutatrice à travers la diode est limitée par le circuit à 10^{-9} sec. Celle-ci, ainsi que d'autres mesures indirectes permettent une évaluation du produit gain-largeur de bande, pour les diodes plus rapides, qui serait de 3 kMc/s.

Zusammenfassung—Die Stromdichte im Tunneleffekt für einen sehr engen p–n-Übergang hängt exponentiell von der reduzierten Effektivmasse der Ladungsträger und der Breite der verbotenen Bandes Energie perre ab. Indiumarsenid scheint deshalb ein geeignets Material zur Herstellung von Tunneldioden hoher Frequenz bei Zimmertemperatur ($m_r \simeq 0{,}05$; $E_g \simeq 0{,}33$ eV). Es gibt bei gegebener Dotierung höhere Frequenzgrenzen und bei gegebenem Produkt aus Verstärkung und Bandbreite eine neidrigere Dotierung (niedrigere Kapazität).

Die Herstellung erfolgt durch Auflegieren von In–Cd- und In–Zn-Pünktchen auf InAs-Kristalle, die mit 2×10^{16} bis 6×10^{18} Te Atomen pro cm^3 dotiert worden waren. Die entstehenden I–V-Kennlinien variieren mit der Dotierung und Temperatur von klassischen p–n-Kurven bis zu Esaki-Kurven, bei denen das Verhältnis von Maximal- zu Minimalstrom 10:1 beträgt. Das beste Verhältnis bei Zimmertemperatur beträgt etwa 2:1.

Gute Elemente mit hoher Stromdichte lassen keine Phononenwirkungen erkennen. Elemente mit geringerer Stromdichte zeigen dagegen zwei Minima der Leitfähigkeit im ansteigenden Teil der Vorwärtskennlinie bei 4°K. Die Elemente grosser Stromdichte zeigen auch eine andere Temperaturabhängigkeit des Spitzenstroms und der Spitzenspannung wie die niedriger Stromdichte. Es liegt deshalb im Bereich der Möglichkeit, dass die Kennlinien der Schwachstromelemente durch einen von Phononen unterstützten Tunneleffekt bestimmt werden, während bei den Starkstromelementen der direkte Tunneleffekt überwiegt.

In den I–V-Kennlinien von Elementen aus schwach dotiertem Material beobachtet man eine interessante Spitzenstruktur. Sie ähnelt der oben erwänten "Phononenkurve", hängt aber von der Oberflächenbeschaffenheit der Diode ab.

Eine Messung der Anstiegszeit eines Stromstosses durch die Diode war durch den Schaltkreis auf 10^{-9} sec begrenzt. Diese und andere indirekte Messungen ergeben einen Wert des Produkts aus Verstärkung und Bandbreite für die Dioden höherer Frequenz von 3 kMHz/sec.

NOTATION

m_r	reduced effective mass	g
E_g	bandgap	eV
W	width of junction	cm
ϵ	dielectric constant	F/cm
e	electronic charge	C
N	doping level	impurities/cm^3
μ_n	n-type mobility	cm^2/V-sec
m_0	electron rest mass	g
R	negative resistance	Ω
r	series resistance	Ω
C	junction capacitance	F
f_{co}	cutoff frequency	c/s
GB	gain-bandwidth	c/s

INTRODUCTION

TUNNEL diodes from III–V semiconductors are of interest for several reasons. Firstly, the wide choice of these materials (as compared to the few elemental semiconductors) offers the possibility of selecting a semiconductor with material constants suited to the particular application of the diode. Secondly, the different band structures of these compounds can be expected to give a variety of interesting electrical characteristics; or in turn, the electrical characteristics can yield new information about these band structures.

We will now give a simple comparison of various semiconductors based on known material constants from the point of view of their application in high-frequency tunnel diodes.

The tunnel current through a very narrow p–n junction in any semiconductor is proportional to the exponential factor

$$\exp[-\alpha(m_r E_g)^{1/2} W] \quad (1)$$

where α is a constant, m_r is the reduced effective mass and E_g the energy gap.* The width of the space-charge region W for a moderately doped symmetrical junction is given by

$$W = \beta[E_g \epsilon/(eN)]^{1/2} \quad (1)$$

where β is a constant, N is the doping level of the

* For lack of better data and for the sake of simplicity, we use the available density-of-states masses and the room-temperature band gaps of the pure materials. For the same reason, the influence of "indirect" tunneling is not considered here.

semiconductor, ϵ its dielectric constant and e the electronic charge. With this we find an expression for the gain-bandwidth product:[2]

$$GB = 1/(2\pi RC) \propto \exp\{-(E_g\gamma)\cdot[\epsilon m_r/(eN)]^{1/2}\} \quad (2)$$

Here γ is a constant, R is the negative resistance and C the capacity of the diode. Equation (2) shows that materials with small bandgap and small effective mass will give the highest gain-bandwidth product for a given doping level.

For the purpose of comparison, we have listed the pertinent material constants of five semiconductors in Table 1. Column 3 is the ratio of the material parameters which appear in the exponential factor in equation (2). It shows InSb and InAs to be the two materials with highest tunneling probability.

Column 4 gives the same information in a more practical way. It gives the doping level N which is necessary to obtain a given factor $E_g(m_r/N)^{1/2}$, i.e. roughly a given GB. (The values N are normalized to give 4×10^{19} cm^{-3} for Ge.) The column shows high speed should be expected on InSb and InAs at low doping. This means that tunnel diodes made from these materials can be expected to have lower capacities (higher impedances) than Ge or GaAs diodes of the same gain-bandwidth product and the same junction area.

where r is the series resistance. The cutoff frequency goes through a maximum as R is decreased by higher doping, since a change in doping changes r and C relatively slowly. The maximum is reached for $R = 2r$, or $f_{co} = 1/(4\pi rC)$. Since $1/r$ for any given geometry is proportional to the mobility μ_n of the majority carriers in the base material*, we have

$$f_{co\ \max} \propto \mu_n \quad (4)$$

These mobility values are listed in the last column of Table 1 for the doping, N of column 4. They show again the superiority of InAs and InSb.

The "fastest" of the five semiconductors listed, InSb, cannot be used at room temperature because of its small bandgap. This has been shown experimentally by BATDORF *et al.*[3] It can be expected, however, that InAs with its bandgap of 0·33 eV will still be extrinsic at room temperature at the higher doping levels, and should exhibit the Esaki effect at room temperature. This latter point has been confirmed by the experimental results described below.

FABRICATION TECHNIQUES

The indium arsenide crystals were doped n-type to the desired concentration (2×10^{16} to 1×10^{19} cm^{-3}) with tellurium or selenium during the crystal-growing process. They usually contained several large single-crystal grains. The crystals were sliced and surface-ground to the desired wafer thickness, usually 10 mils, diced by scribing with a sapphire point, and etched in hot dilute HNO_3.

Indium dots containing about 10 per cent cadmium

Table 1. Material constants pertinent to tunneling

	E_g (eV)	m_r/m_0	$E_g(m_r/m_0)^{1/2}$ (eV)	N (cm^{-3})	μ_n (cm^2/V sec)
Ge	0·67	0·44	0·44	4×10^{19}	220
Si	1·11	0·78	0·98	2×10^{20}	40
GaAs	1·35	0·13	0·49	5×10^{19}	800?
InAs	0·33	0·051	0·075	1×10^{18}	10,000
InSb	0·18	0·028	0·030	2×10^{17}	45,000 (at 77°K)

Obviously, all these values are only approximate because of the uncertainties in E_g and m_r for degenerate material. But they serve their purpose for the above comparison of the listed semiconductor materials.

Another characteristic frequency is the cutoff frequency[2]

$$f_{co} = [(R/r)-1]^{1/2}/(2\pi RC) \quad (3)$$

* We assume the more favourable case where the main contribution to r is in n-type material.

or 5–20 per cent zinc were alloyed to the surface of the InAs pellets in a strip-heater furnace with a reducing atmosphere. Better uniformity of wetting was obtained by the use of an acid flux. The alloying temperature ranged from 380 to 480°C, and this temperature was usually kept for 1 min. Slow cooling (about 1°C/sec) and heat supplied from the top by a second heater strip helped to get a good regrowth layer.

a considerable change in their characteristics upon exposure to water vapor or when taken out of a cold bath in the moist room air. Diodes so treated can be restored by boiling in acetone.

I–V CHARACTERISTICS

Figs. 1–5 show the *I–V* characteristics at several different temperatures of typical units made from material doped from 2×10^{16} to

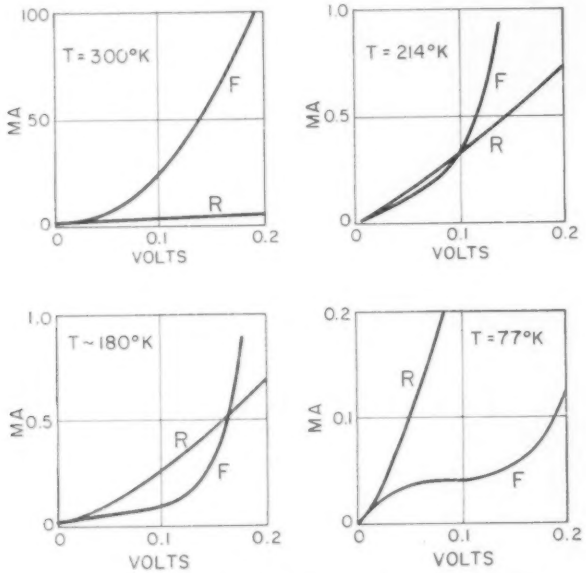


FIG. 1. Diode no. 112 : $N = 2\times10^{16}\,\text{cm}^{-3}$; $A = 1{\cdot}6\times10^{-3}\,\text{cm}^2$. (Note different current scales.)

The base connection was made by soldering the pellet with tin–lead eutectic and acid flux to a nickel base tab or into a gold-plated Kovar enclosure. Tin and lead are both donors in InAs[4] and make a good ohmic contact to the *n*-type wafer. Finally a platinum wire or a gold-plated Kovar finger was fused to the dot, again with the help of flux.

Most of the units could be improved by etching with a mixture of 6 volume parts concentrated HNO_3, 3 parts acetic acid and 1 part HF; one drop of saturated KI solution was added to 10 ml of this. Diodes made from material containing more than 3×10^{18} donors/cm^3 are quite insensitive to surface effects. However, lower doped units show $5{\cdot}8\times10^{18}\,\text{cm}^{-3}$. In Fig. 1 ($N = 2\times10^{16}$) the room-temperature curve is a normal *p–n* characteristic. At 214°K the reverse current, determined by Zener tunneling, has not changed much. However, the forward current, which is mainly injection, has become lower than the reverse current for low voltages. At 77°K the injection current is low enough, so that the tunnel current shows up in the forward direction as a shoulder. We have a so-called tunnel rectifier.

The tunneling component is bigger in units made from higher doped material [see equation (1)]. This is apparent in Fig. 2 ($N = 5{\cdot}4\times10^{17}$), where we already have a negative resistance at 77°K. With increasing doping level the tunnel

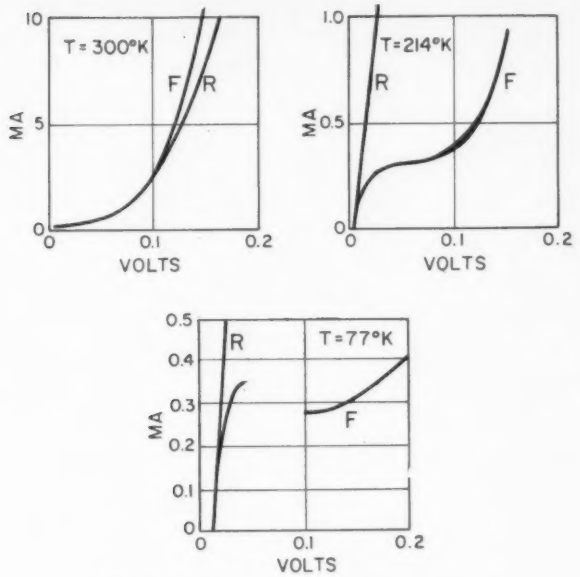


FIG. 2. Diode no. 87 : $N = 5{\cdot}4 \times 10^{17}$ cm^{-3}; $A = 2{\cdot}5 \times 10^{-3}$ cm^2.

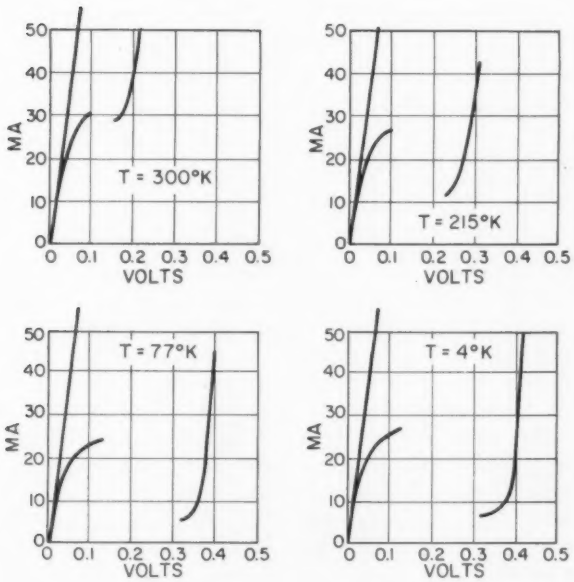


FIG. 3. Diode no. 649 : $N = 3{\cdot}1 \times 10^{18}$ cm^{-3}; $A = 6 \times 10^{-5}$ cm^2.

current dominates more and more, and the negative resistance occurs at higher temperatures as shown in Figs. 3–5 ($N = 3{\cdot}1 \times 10^{18}$, $4{\cdot}8 \times 10^{18}$ and $5{\cdot}8 \times 10^{18}$). It can be seen that at room temperature the valleys are quite narrow and steep indicating that the classical current obscures the excess current. Only at low temperatures do some of the curves show a wider trough indicating the presence of an excess current.

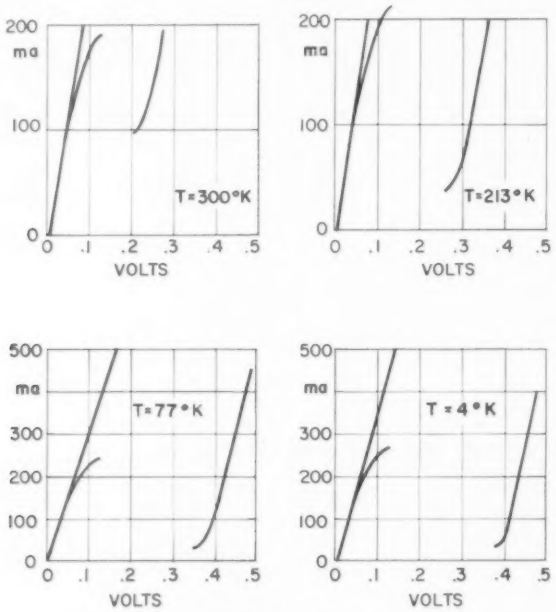


FIG. 4. Diode no. 199 : $N = 4{\cdot}8 \times 10^{18}$ cm^{-3}; $A = 1{\cdot}6 \times 10^{-4}$ cm^2.

The particular unit shown in Fig. 5 does not have quite as high a peak to valley ratio as the slightly lower-doped unit in Fig. 4. This is typical for all the diodes made from higher-doped material, and it is not clear what causes it. Experiments with Ge indicate that this might be due to crystal imperfections.[5] Fig. 6 shows the peak current density versus doping for a number of units.

A close examination at 4°K of the I–V characteristics of many units shows cusps similar to those described by HOLONYAK *et al.*[6] Figs. 7 and 8 are examples of these; two conductance minima are easily visible on the rising part of the forward characteristic. Table 2 gives the voltages (corrected for series resistance obtained from the high-current region of the reverse characteristic) at which these minima occur for units with various dopings. The values are in fair agreement with those reported by HALL *et al.*[7] who attribute these minima to tunneling processes involving one and two longitudinal optical phonons, respectively.

These cusps, however, do not appear on all diodes. In particular, the units with high current densities and good current ratios usually have smoothly varying characteristics. This indicates that in the high current units the direct tunneling component is dominant over the phonon-assisted current. The fact that the tunneling processes dominating the two types of diodes are different is also supported by the temperature dependence of the peak currents. This is demonstrated in Fig. 9, where the peak currents of smooth units (dotted curves) and of units with cusps (solid curves) are plotted versus the temperature. The strikingly different behavior is immediately evident.

The strong increase of the tunnel current from

Table 2. Voltages at which conductance minima occur

N	Diode No.	1st minimum (mV)	2nd minimum (mV)	Figure no.
$5{\cdot}8\times10^{18}$	433	25	70	8
$5{\cdot}8\times10^{18}$	529	22	75	—
$3{\cdot}1\times10^{18}$	645	27	65	7
$3{\cdot}1\times10^{18}$	649	25	55	—
$5{\cdot}4\times10^{17}$	87	25	—	11(b)
$5{\cdot}4\times10^{17}$	87	23	54	12
$5{\cdot}4\times10^{17}$	87	—	70	11(c)
$5{\cdot}4\times10^{17}$	87	—	60	11(d)
$5{\cdot}4\times10^{17}$	87	25	—	11(f)
$5{\cdot}4\times10^{17}$	87	—	60	11(g)

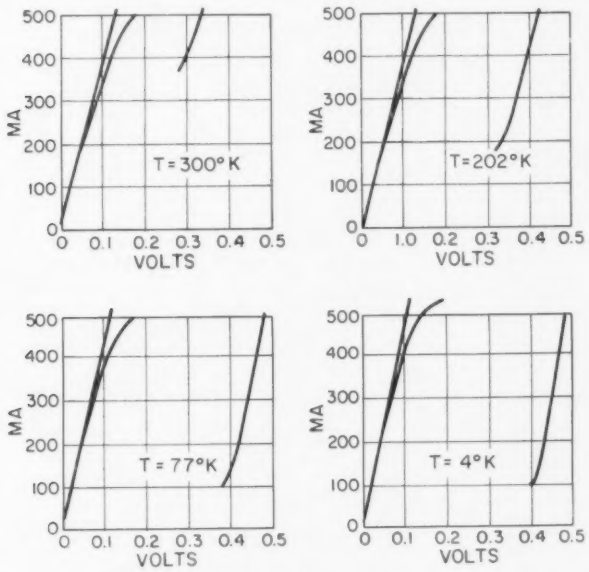


FIG. 5. Diode no. 488 : $N = 5{\cdot}8\times10^{18}$ cm^{-3}; $A = 10^{-4}$ cm^2.

200 to 300°K can be ascribed to the increase of phonon density, which causes an increased stimulated emission and an increased absorption of phonons. This mechanism for current increase versus temperature should not be present in the units in which the direct tunneling dominates. In addition, the peak voltages of the two types have different temperature coefficients, as can be seen in Fig. 10.

It has been mentioned above that diodes made from lower-doped InAs are sensitive to changes of surface conditions. We discovered that in some cases cusps can be introduced or removed by suitable surface treatments. The evidence is given in the curve tracer photographs shown in Fig. 11. The diode is the lightly doped unit of Fig. 2. Fig. 11(a) gives the characteristic of the freshly etched unit after boiling in acetone and immersion in

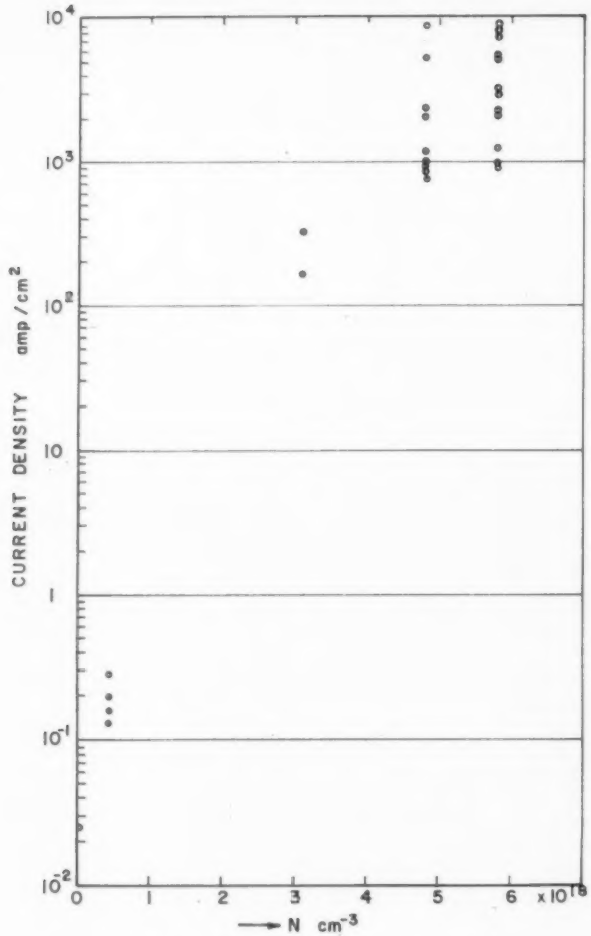


FIG. 6. Peak current density vs. doping.

liquid nitrogen. The diode was then taken out of the liquid nitrogen, allowed to warm up and dry and was immersed in liquid helium, Fig. 11(b). Note that the cusp current rises above the 77°K curve. After warming up in air and again immersing in helium the unit has lost its negative resistance but the cusps have become bigger, as shown in Fig. 12 (note the different scales: 0·1 mA/division and 10 mV/division). After warming up, the characteristic in liquid nitrogen shows a higher current and cusp structure, Fig. 11(c) (the interruptions and double trace at the right side were shown to be oscillations). Short boiling in acetone produced the 77°K curve of Fig. 11(d). A second boiling step in acetone returned the diode to its initial condition as shown in Fig. 11(e) at 77°K and Fig. 11(f) at 4°K. These are the same as Figs. 11(a) and 11(b), respectively. After warming up, immersing into a dry-ice–acetone bath, warming up and immersing in liquid nitrogen again, it shows the 77°K curve of Fig. 11(g), which is similar to Fig. 11(d).

It seems to be reasonable to assume that these changes are due to moisture condensed at the surface, which could be removed effectively by boiling in acetone. The significance of this experi-

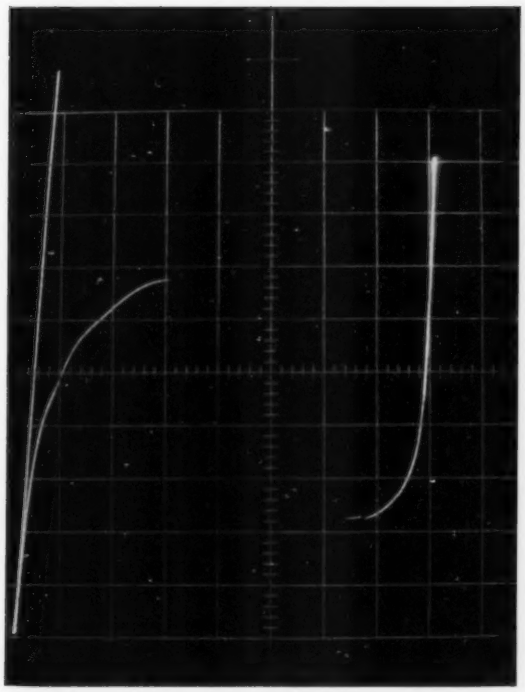


FIG. 7. Diode no. 645 ($N = 3{\cdot}1 \times 10^{18}$ cm^{-3}) at 4°K. Vertical scale 2 mA/div, horizontal scale 50 mV/div.

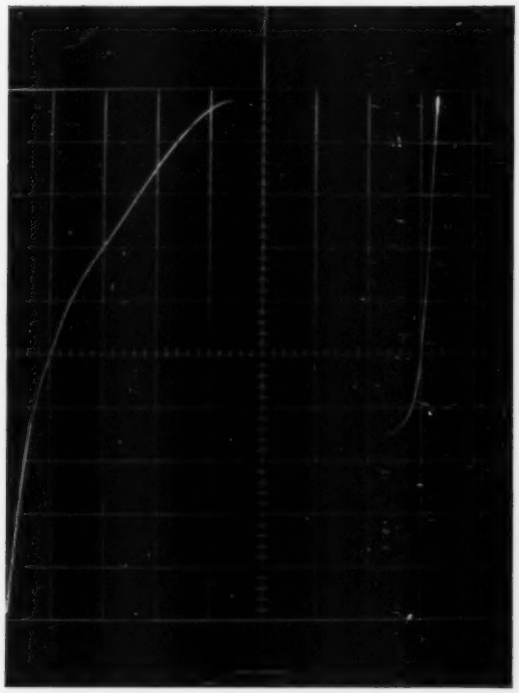


FIG. 8. Diode no. 433 ($N = 5{\cdot}8 \times 10^{18}$ cm^{-3}) at 4°K. Vertical scale 10 mA/div, horizontal scale 50 mV/div.

[*facing p.* 140

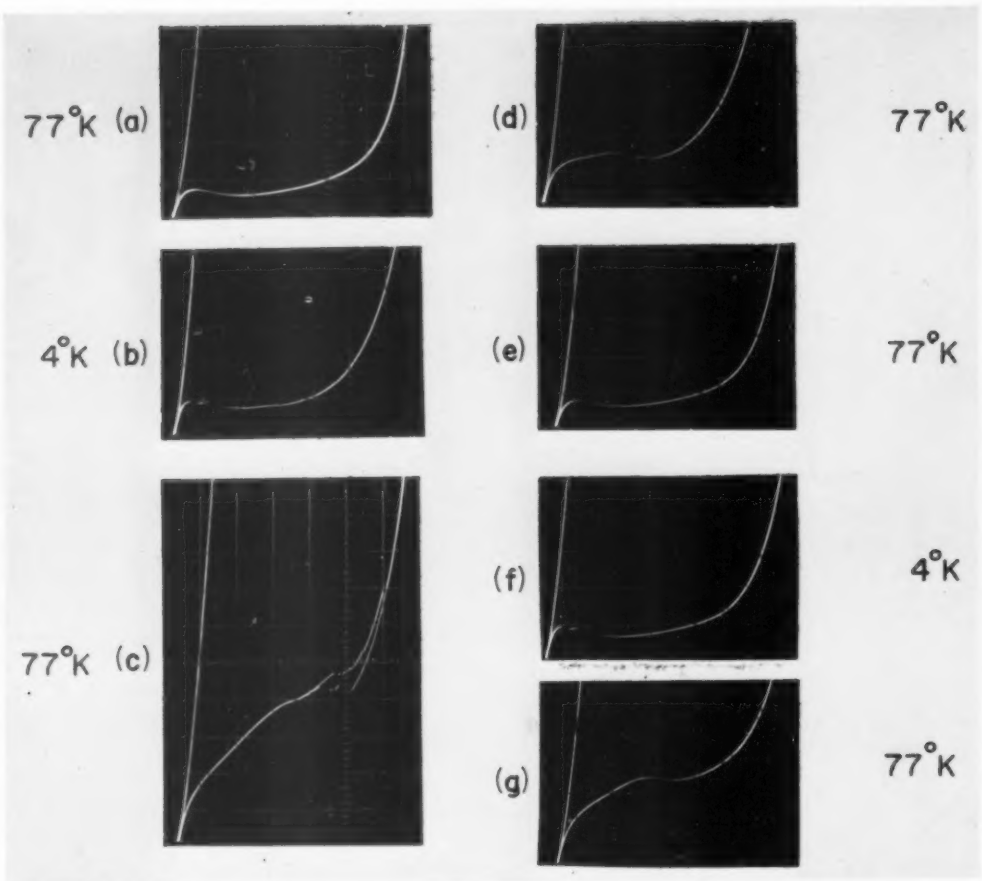


FIG. 11. Characteristics of diode no. 87 ($N = 5{\cdot}4 \times 10^{17}$ cm^{-3}) at different surface conditions. Horizontal scale 50 mV/div, vertical scale 0·5 mA/div.

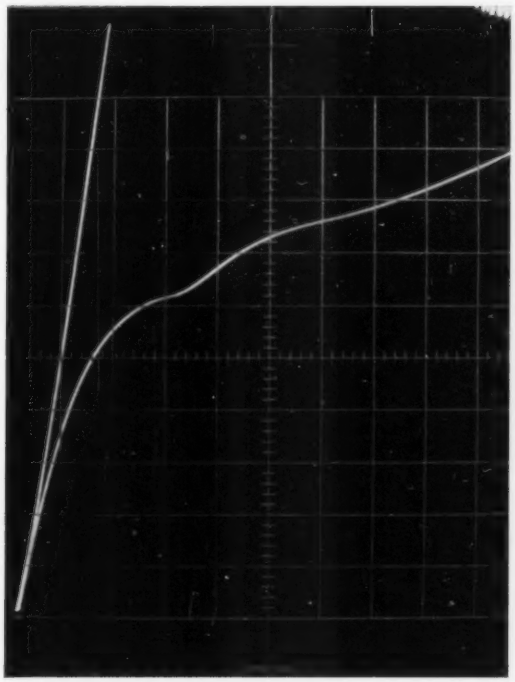


FIG. 12. Horizontal scale 10 mV div, vertical scale 0·1 mA/div.

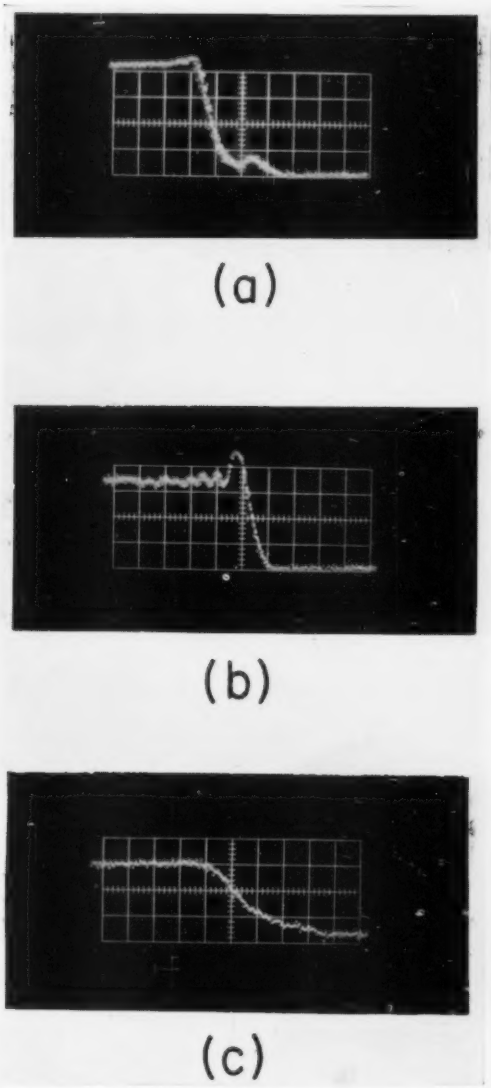


FIG. 14. (a) Pulse rise time of diode no. 488, (b) of same circuit with 2-Ω resistor instead of the diode and (c) of a slower diode, no. 377; horizontal scale : 1 mμsec/div.

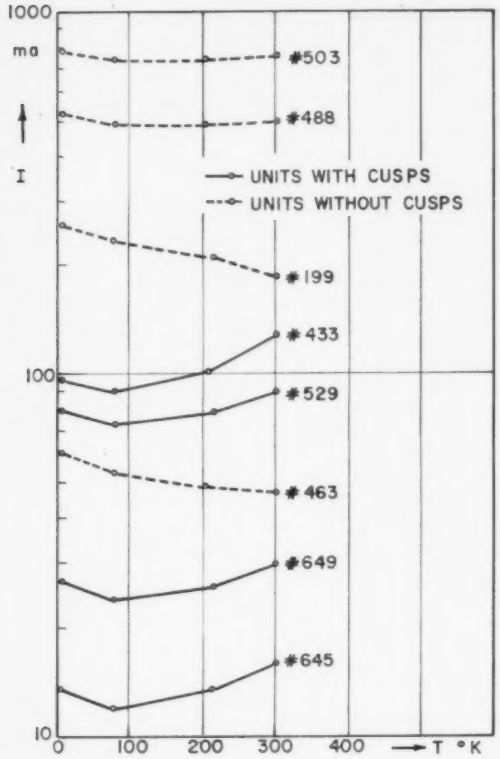


FIG. 9. Peak current as function of temperature.

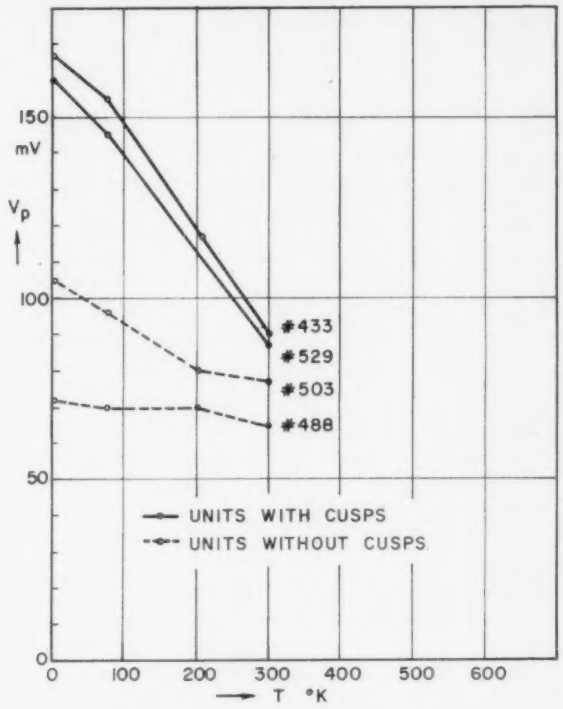


FIG. 10. Peak voltage as function of temperature.

ment is that it shows the cusps can be caused by surface phenomena. It is interesting to note that the positions of the minima in Figs. 11 and 12 agree with the ones found on higher-doped units, which are listed in Table 2.

PRELIMINARY RESULTS ON THE HIGH-FREQUENCY PROPERTIES

All of the InAs diodes fabricated so far have areas larger than 10^{-4} cm^2, and the units made from the higher-doped material, i.e. the high gain-bandwidth units, have tunnel currents of over 100 mA. For this reason, we have not been able to measure the capacities C and the negative resistances R directly. However, an estimate can be made based on the approximate area measurement and the doping level and by extrapolation measurements on lower-current units. This gives, for the unit no. 488 of Fig. 5, $R = 0{\cdot}5\ \Omega$ and $C = 100\ \mu\mu$F, and hence GB $\simeq$ 3 kMc/s.

On this same diode the rise time of a switching pulse was checked. Fig. 13 shows the circuit used and Fig. 14(a) shows the trace on the sampling scope. Figure 14(b) gives the same pulse with a 2-Ω resistor in place of the diode. One can see that there is no difference in rise time between the two. Hence the diode rise time is at least three times better than the 1 mμsec rise time of the circuit and scope. Hence, a reasonable guess for the diode is 0·3 mμsec, which is in agreement with the estimate of GB given above. Fig. 14(c)

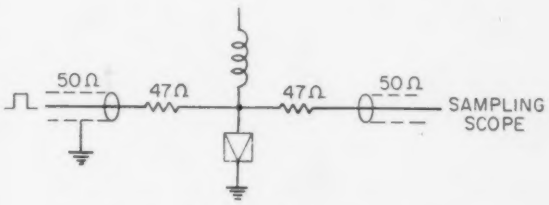


FIG. 13. Circuit for rise-time measurement.

gives the same pulse trace with a slower diode having a negative resistance of 5 Ω. The measured pulse rise time of 3 mμsec checks with our estimate.

CONCLUSION

An analytical comparison of the tunneling probability of various semiconductors shows that InAs offers good possibilities for tunnel diodes suitable for centimeter wavelengths and operation at room temperature. Tunnel diodes have been made which have current ratios of 2:1 at room temperature and 8:1 at 80°K. Diodes made with a base material with 5×10^{18} electrons/cm^3 have a figure of merit of 3 kMc/s.

In many of the diodes there is a cusp structure in the I–V characteristic at 4°K. In some of the low-doped tunnel diodes, this seems associated with phonon-assisted transitions, although in others it is definitely associated with surface effects. Correlations have been made between the temperature dependence of the d.c. characteristic and the nature of the cusp structure.

Further studies on InAs tunnel diodes would be interesting for both device development and scientific reasons. Higher-doped units should be easier to use in microwave circuits than their germanium counterparts because of their lower specific capacitance. More work is needed before we have an understanding of the peculiar effects of surface treatment.

Acknowledgements—It is a pleasure to give acknowledgement to many members of this laboratory. Especially, we want to thank H. S. Sommers, Jr., H. Nelson and C. W. Mueller for many helpful discussions, R. M. Williams for discussions and help with the manuscript and R. C. Conti for experimental assistance; the pulse measurement was done by M. H. Lewin.

REFERENCES

1. E. O. Kane, *J. Phys. Chem. Solids* **12,** 181 (1959).
2. H. S. Sommers, Jr., *Proc. I.R.E.* **47,** 1201 (1959).
3. R. L. Batdorf, G. C. Dacey, R. L. Wallace and D. J. Walsh, *J. Appl. Phys.* **31,** 613 (1960).
4. E. Schillmann, *Z. Naturf.* **11**a, 463 (1946).
5. H. Nelson. Private communication.
6. N. Holonyak, Jr., I. A. Lesk, R. N. Hall, J. J. Tiemann and H. Ehrenreich, *Phys. Rev. Letters* **3,** 167 (1960).
7. R. N. Hall, J. H. Racette and H. Ehrenreich, *Phys. Rev. Letters* **4,** 456 (1960).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 143–148. Printed in Great Britain

NOTES ON THE THEORY OF FOUR-LAYER SEMICONDUCTOR SWITCHES

A. K. JONSCHER

Research Laboratories of The General Electric Company Limited, Wembley, England

(*Received* 6 *October* 1960)

Abstract—Starting from the accepted theory of four-layer devices, a discussion is developed of the effects of transverse base resistance on three-terminal operation of *p–n–p–n* switches. The treatment is then extended to transient response to two- and three-terminal on- and off-switching.

Résumé—En partant de la théorie acceptée des éléments à quatre couches, on développe une discussion des effets de la résistance transversale de base sur l'opération à trois bornes des commutateurs *p–n–p–n*. On élargit ensuite la description de façon à y inclure la réponse transitoire des commutateurs à deux et trois bornes.

Zusammenfassung—Von der etablierten Theorie der Vierzonengeräte ausgehend, werden die Wirkungen des transversalen Basiswiderstands auf die Arbeitsweise eines *p–n–p–n*-Schalters mit drei Kontakten erörtert. Es folgt eine Behandlung der Einwirkung des Trägerdurchgangs auf die Geschwindigkeit des Ein- und Abschaltens bei zwei oder drei Kontakten.

1. INTRODUCTION

MANY aspects of the theory of operation of four-layer (*p–n–p–n*) devices, both two- and three-terminal, are now well established and commonly known. There are some aspects, however, especially the transient response and the effect of transverse base resistance on three-terminal operation, which are not sufficiently appreciated. The present paper sets out to deal with a few problems falling in this category. Its point of departure is standard theory as developed in the literature[1] and, in particular, it follows the treatment and notations employed by the author elsewhere.[2]

No attempt is made here at a rigorous mathematical treatment since it is felt that the complexity of the problem is too great for this to be profitable. Instead, semi-qualitative discussion is given of the various phenomena involved.

2. THE BASIC EQUATIONS

It is proposed to analyse the operation of a four-layer structure shown schematically in Fig. 1. It consists of three *p–n* junctions 1, 2 and 3, separating four regions denoted p_1, n_1, p_2 and n_2. Ohmic connexions are provided to the outer regions p_1 and n_2, referred to as the emitters, and also to one of the two internal base regions, n_1 in the diagram. The two base regions are characterized by four

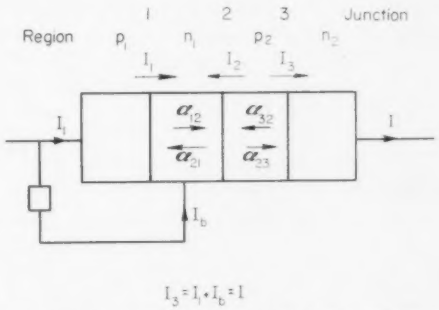


FIG. 1. The three-terminal four-layer structure.

current-gain parameters, α^0_{mn}, m and n taking the values 1, 2 and 3. These gain parameters are defined in accordance with standard transistor theory with the junction m emitting and junction n short-circuited.* When appreciable reverse bias is applied

* In transient operation the large-signal current gains are applicable.

to the collecting junction n, a current multiplication factor α^* is introduced to give the effective gain

$$\alpha_{mn} = \alpha_n^* \alpha_{mn}^0 \tag{1}$$

The current in any of the junctions is defined in terms of a saturation current I_{nS} and a bias parameter ν_n given by

$$\nu_n = \exp\left(\frac{eV_n}{kT}\right) - 1 \tag{2}$$

where V_n is the bias reckoned positive in the forward direction. Thus

$$I_n = I_{nS}\nu_n \tag{3}$$

The equations governing the behaviour of the structure may be written in the following form [cf. Ref. 2, equation (5.33)]:

$$\left.\begin{aligned} I - I_b &= I_{1S}\nu_1 - \alpha_{21} I_{2S}\nu_2 \\ I - I_{20} &= \alpha_{12} I_{1S}\nu_1 - \alpha_2^* I_{2S}\nu_2 + \alpha_{32} I_{3S}\nu_3 \\ I &= -\alpha_{23} I_{2S}\nu_2 + I_{3S}\nu_3 \end{aligned}\right\} \tag{4}$$

Here $I_{20} = I_{20}(V_2)$ is the voltage-dependent reverse excess leakage current on the middle junction. This accounts for any leakage current in excess of the value given by (3). The corresponding terms have been omitted from the other equations, since junctions 1 and 3 operate in the forward direction. For the same reason the multiplication coefficients α_1^* and α_3^* have been equated to unity.

The overall potential difference across the device is

$$V = V_1 - V_2 + V_3 \tag{5}$$

and is positive when the p_1 region is positive. For positive applied bias and positive current I, junctions 1 and 3 operate in the forward direction, junction 2 has a reverse bias in the pre-breakdown condition and a forward bias in the post-breakdown condition.

The system of equations (4) can be solved for the bias parameters ν_n and the solution relevant to the following discussion is that for the middle junction

$$\nu_2 = -[I(1 - \alpha_{12} - \alpha_{32}) - I_{20} + I_b\alpha_{12}]/(I_{2S}h) \tag{6}$$

where

$$h = \alpha_2^* - \alpha_{12}\alpha_{21} - \alpha_{23}\alpha_{32} = \alpha_2^* h^0 \tag{7}$$

is an essentially positive parameter, both with and without multiplication in the middle junction. h^0 is the low-voltage value of h.

3. STEADY-STATE OPERATION

3.1. *Pre-breakdown conditions*

Under pre-breakdown conditions of a normal device the sum of inward alphas is less than unity and therefore the middle junction is biased in the reverse. Thus $\nu_2 = -1$ and the current is limited to a value given by

$$I = \frac{hI_{2S} + I_{20} - I_b\alpha_{12}}{1 - \alpha_2^*(\alpha_{12}^0 + \alpha_{32}^0)} \tag{8}$$

The alphas depend upon current in a manner represented schematically in Fig. 2. Their sum

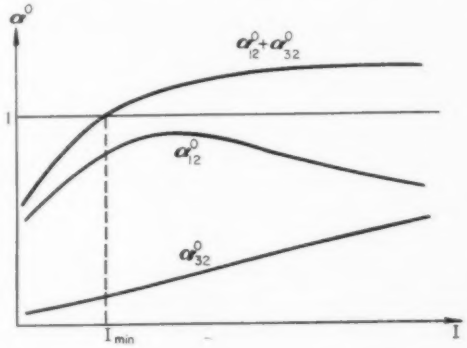


FIG. 2. Current dependence of low-voltage gain parameters α_{12}^0 and α_{32}^0.

reaches unity at some current $I = I_{\min}$, and below this value equation (8) applies.

Two points should be borne in mind. Firstly, each alpha depends on current flowing across the respective emitter, i.e. α_{12} depends on $I_1 = I - I_b$, while α_{32} depends on I. This is relevant in the presence of base current. Secondly, the alphas may be affected by non-uniform distribution of current density. Thus, in the absence of base current the alphas are higher than the corresponding values measured on a transistor with a base connexion, owing to the absence of non-uniform distribution of emission density.

3.2. *Breakdown from low- into high-conductance state*

As the overall voltage V across the device increases, V_2 becomes more negative. This may have

a twofold effect. The leakage current $I_{20}(V_2)$ may increase sufficiently to cause a rise of I beyond $I_{\min}$, thus causing the denominator to change sign. This type of breakdown is very temperature sensitive.

Alternatively, with sufficiently low leakage current, the increase of voltage V_2 may cause the multiplication coefficient α_2^* to rise to the extent that the denominator vanishes. This mode of avalanche breakdown is less temperature sensitive and therefore more desirable than the former mode. In practice both mechanisms are present simultaneously, especially in view of the fact that I_{20} is subject to the same multiplication α_2^*, apart from any other voltage dependence which it may have.

The break-over current I_{BO}, which is the maximum current the device can sustain with reverse bias on junction 2, is given by the condition

$$\alpha_{12}+\alpha_{32} = \alpha_2^*(\alpha_{12}^0+\alpha_{32}^0) = 1$$

This current may be less than $I_{\min}$, except where the leakage current I_{20} is large and concentrated near the surface. The gain parameters α may be expected to be lower there on account of surface recombination, and the break-over current may then be higher than for uniform distribution of current across the emitter face. In fact it is observed experimentally that, in "leaky" devices, I_{BO} can be higher than I_H, the holding or minimum sustaining current, discussed in Section 3.3.[3]

An alternative method of triggering consists in the application of a negative base current I_b. This causes an increase of I and consequently a decrease of the denominator in (8) until breakdown ensues. The current I_b is the more effective the higher α_{12} and the smaller the value of current $I_{\min}$ at which $\alpha_{12}+\alpha_{32} = 1$.

The negative base current required for triggering has the same polarity as the base current supplying recombination and emitter inefficiency losses in an ordinary transistor. This current gives rise, therefore, to the same kind of transverse potential drop in the base region and decrease of emission with distance away from the base terminal. It would appear, however, that to trigger the device it is not necessary to induce enhanced emission across the whole of the emitter face. Once enhanced emission has been brought about locally, breakdown will ensue, even if with some delay.

3.3. *Post-breakdown operation*

The high-conduction condition after breakdown is characterized by forward bias on all three junctions, the middle junction operating in opposition to the outer two. The alphas now satisfy the relation

$$\alpha_{12}^0+\alpha_{32}^0 > 1 \tag{9}$$

the current is high, v_2 is positive and may be large. Setting $I_{20} = 0$ for the forward-biased junction, one obtains from equation (6)

$$v_2 = [I(\alpha_{12}^0+\alpha_{32}^0-1)-I_b\alpha_{12}^0]/I_{2S}h^0 \tag{10}$$

The opposing or stalling action of the middle junction is caused by the requirement of continuity of flow. The higher the quantity $\alpha_{12}^0+\alpha_{32}^0-1$ the more stalling action is required.

The minimum sustaining current is that current below which the high-current regime can no longer be maintained. This is clearly the current $I_{\min}$, defined in Fig. 2, at which $\alpha_{12}^0+\alpha_{32}^0$ passes through unity. The device can be turned off by reducing the current below this value for a sufficient length of time (see Section 4.3).

An alternative method of turning off the device consists in the application of a positive current to the base. A base current

$$I_b > I(\alpha_{12}^0+\alpha_{32}^0-1)/\alpha_{12}^0 \tag{11}$$

causes the bias V_2 to become negative and reduces the bias on junction 1. It is evident that to minimize the base current required for turning off the device the base region should have a high alpha and the sum of alphas should exceed unity by as small an amount as possible at the operating current.

The presence of the transverse base resistance makes the turning-off operation more difficult than it would have been with zero base resistance. Unlike in turning on, where a local enhancement of emission leads to breakdown, for turning off it is essential that condition (11) be realized over the entire emitting area. The resistance of a base region having a reasonably good alpha is likely to be high, especially so since the base region next to the base terminal operates under conditions of no

injection, that is with unmodulated conductance. This may be seen from the diagram in Fig. 3, where the solid bands correspond to some point in the interior of the base region, while the dotted line shows the n-type base nearer the base terminal.

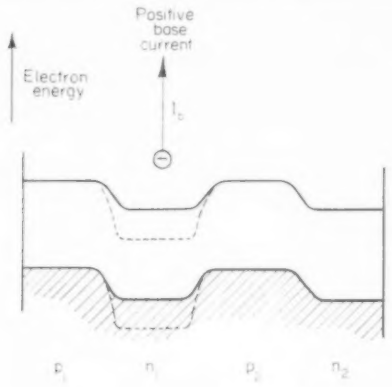


FIG. 3. Reverse biasing of the emitter in the process of turning off as a result of transverse potential drop in the base region.

The difference is due to the potential drop in the base region caused by the flow of heavy base current. The implication is that the left-hand emitter is cut off over a substantial part of its area and the base current for turning off has to be even higher than the value given by equation (11). The design of controlled rectifiers should therefore take account of the base resistance, at least as much as the design of power transistors, if satisfactory turning off via the base terminal is to be achieved.

4. TRANSIENT RESPONSE

The foregoing considerations applying to steady-state or slowly variable conditions need considerable modification in the case of fast transients. Two kinds of characteristic time constants may be defined for the device. In the first category are the lifetimes of injected excess carriers in the inner base regions, denoted by τ_1 and τ_2 for n_1 and p_2 respectively. The second category is the diffusive transit times, defined as

$$t_{t1,2} = \frac{W_{1,2}^2}{2D_{1,2}}$$

where W and D are the respective base widths and diffusion constants. The response to transients fast compared to the longest of these time constants may differ from steady-state response.

A further time dependence is introduced by the necessity to supply or remove definite amounts of charge on going over from one steady-state condition to the other. This involves time through the magnitude of current determined by the external circuit and by the condition of the device.

4.1. *Two-terminal breakdown*

Rapid application of a positive bias, or increase of a standing bias, to the device in its low-conductance condition results in a reverse bias on the middle junction.

A definite amount of charge Q_c must be shifted from the space-charge region of this junction, by injection into the outer regions. This, however, causes injection from the outer emitters of minority charge into the base regions, Fig. 4. Any charge

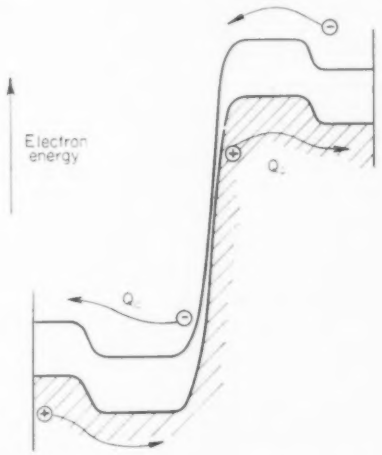


FIG. 4. Transient application of positive bias to the structure.

injected into these regions and, if sufficient time is available, transported across and collected in the opposite regions, contributes to "charge gain". This means that more charge has in effect to be passed through the device than the amount strictly required for charging the middle junction. The faster the rise time the higher the currents involved and, consequently, the higher the inward alphas. This leads to the reduction of breakdown voltage

under impulse triggering and this reduction would be minimized if the emitter efficiencies and alphas for inward injection on junctions 1 and 3 were kept low.

The charging process, including any unnecessary transport outlined above, takes a time depending mainly on the external circuit. Simultaneously, the excess densities are being transported across the two base regions and this process is completed after the longer of the two times t_{t1} and t_{t2}. Before this instant the alphas have no meaning since no collection takes place. If at that instant the sum of alphas already exceeds unity, the process of collapse of the reverse bias sets in. The corresponding re-charging of the space-charge region is done at the expense of the main current but only a fraction of it can take part in the process. It is that part which can be diverted from the through-flow without causing the sum of alphas to fall below unity. To that extent again, therefore, the re-charging time depends partly on the external circuit and partly on the amount by which $\alpha_{12}^0+\alpha_{32}^0$ exceeds unity.

In conclusion, this mode of breakdown takes a time equal to the longer of the two transit times, t_{t1} or t_{t2}, plus such time as may be required for the charging and re-charging of junction capacitance.

4.2. *Breakdown with base triggering*

With reference to Fig. 5, consider the sudden application of a negative base current to the device. The inflowing electrons are injected into the p-region, with consequent forward bias on junction 1 and flow of holes into the base region. These holes arrive in the p_2 region with a delay t_{t1} and cause injection into that region from junction 3. After a further interval t_{t2}, the electrons arrive in n_1 and the sum of alphas may then exceed unity. The ensuing process of re-charging of junction 2 follows similar lines to that described in Section 4.1, except that the supply of electrons is obtained from the base lead and not necessarily across p_2 by collection.

A further inevitable delay may be involved in this process if the base resistance causes localized breakdown only near the base terminal, as explained in Section 3.2. The delay due to transverse spreading is difficult to estimate but it might be appreciable, since an effective limitation of current is involved.

The base terminal triggering, therefore, entails a delay equal to the sum of transit times $t_{t1}+t_{t2}$ plus the re-charging time of junction capacitance plus

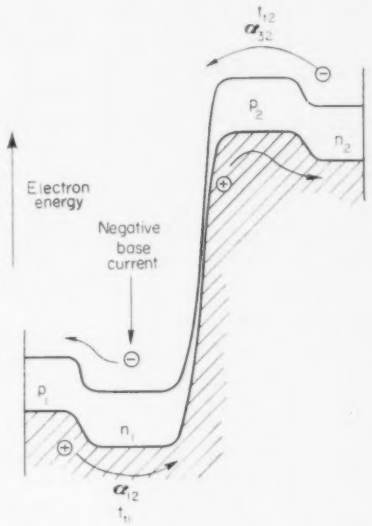


FIG. 5. Triggering by the base pulse.

any delay that may be due to incomplete breakdown across the base region.

4.3. *Turn-off speed*

The process of turning off the device amounts essentially to the disposal of the injected excess carrier densities and the subsequent charging of the space-charge capacitance. A treatment of two-terminal turning-off operation in terms of charge control analysis has been given by BAKER *et al.*[4]

The disposal of excess charge can be effected in two ways—by recombination and by charge withdrawal, depending upon the manner of turning off. Where the current is merely reduced below its sustaining value I_{min}, recombination is the dominant process. Charge withdrawal enters in the case of reversal of current flow in the device. It should be borne in mind, however, that the withdrawal of charge from a four-layer structure still entails some injection and diffusive transport, unlike an ordinary diode. With reference to Fig. 3, the majority charge in, say, the p_2 region can escape only into n_1 when junction 3 is reverse biased. The same applies to

the region n_1 unless the base current is allowed to flow.

Charge withdrawal by reversal of current flow involves, of course, the usual transient current peak. If this is limited by external resistance the whole process takes a longer time, since a definite amount of stored charge is involved. In the case where the current is severely restricted by external circuit the recombination process takes over.

The conclusion is that, to minimize the recovery time and the amount of charge transported on turning off, it is necessary to reduce the widths of the inner base regions and to reduce the lifetime throughout the device. Fortunately, these requirements are compatible with the need to maintain reasonable values of alphas.

In the case of triode design a compromise may have to be reached, however, on the question of base resistance which would increase with decreasing width.

There is scope in the design of fast switches for reduction of lifetime by deliberate introduction of "lifetime killers". Whether the geometry could be modified so as to make better use of surface recombination remains an open question.

With regard to the process of charging the middle junction capacitance, the charging time depends upon the amount of charge, which involves the "charge gain" discussed in Section 4.1, and upon the current which is allowed to flow in the system.

5. CONCLUSIONS

It is concluded that the base resistance may have a delaying influence on the speed of turning on and may make the turning-off operation more difficult. The speed of response depends critically on the widths of the inner regions, subject to the reservation about the base resistance in three-terminal structures, and also on the lifetime in the device. The use of lifetime killers and possibly of surface recombination is recommended. It is suggested that for easy turning off the sum of alphas should not greatly exceed unity.

Premature two-terminal triggering by a fast transient would likewise be reduced by a decrease of the sum of alphas.

The speed of response is determined jointly by lifetime, transit times in the inner base regions and by the magnitude of circuit-determined charging currents. The two base regions should not have widely differing widths, and it would be preferable if the two inward alphas were approximately equal rather than very different.

Acknowledgement—The author is indebted to Mr. R. D. KNOTT for helpful comments and discussion.

REFERENCES

1. J. L. MOLL, M. TANENBAUM, J. M. GOLDEY and N. HOLONYAK, *Proc. I.R.E.* **44**, 1174 (1956).
2. A. K. JONSCHER, *Principles of Semiconductor Device Operation* Section 5.6. Bell, London (1960).
3. R. D. KNOTT. Private communication.
4. A. N. BAKER, J. M. GOLDEY and I. M. ROSS, *I.R.E. Wescon Convention Record*, Pt. 3, 43 (1959).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 149–156. Printed in Great Britain

PHOTORECTIFIER BASED ON A COMBINATION OF A PHOTOCONDUCTOR AND AN ELECTRET*

J. G. VAN SANTEN and G. DIEMER

Philips Research Laboratories, N.V. Philips' Gloeilampenfabrieken Eindhoven, Netherlands

(*Received* 4 *October* 1960)

Abstract—Photoconductive layers of CdS powder embedded in an electret (glass–enamel) can be formed to become rectifying by the combination of heating and the passage of a d.c. current. The mechanism of the rectification is discussed with reference to the mechanism of photocurrents through powdered CdS layers. A cross-bar system based on this new type of photorectifier has been constructed.

Résumé—Des couches de conducteur photoélectrique de CdS en poudre, encastrées dans un électret (verre–emaillé), peuvent devenir des redresseurs en les soumettant à un traitement de chauffage et de passage d'un courant continu. On décrit le mécanisme de redressement en référence au mécanisme des courants photoélectriques traversant les couches CdS en poudre. Un modèle à traverse basé sur ce nouveau type de redresseur photoélectrique a été construit.

Zusammenfassung—Photoleitende Schichten von CdS-Pulver, das in Glass–Email (Elektret) eingebettet ist, lässt sich durch die Kombination von Erhitzen und dem Durchgang eines Gleichstroms zu einem Gleichrichter formieren. Der Mechanismus der Gleichrichterwirkung wird unter Bezugnahme auf den Mechanismus von Photoströmen durch Schichten von CdS-Pulver erörtert. Mit diesem neuen Typ von Photogleichrichter wurde ein Crossbarsystem konstruiert.

INTRODUCTION

IT IS well known that in semiconductors, rectification can be obtained in various ways. It has long been realized that in order to produce rectification the parts of the semiconductor near to the electrodes must have different conductivities. Rectification can thus be due to (1) a different size of the electrodes, (2) to a difference in the materials used for the two electrodes or (3) to a variation in bulk properties throughout the inter-electrode region. Photorectifiers can be based upon the same principles; moreover in photoconductors the gradient in bulk properties can be brought about by a variation of excitation density throughout the inter-electrode region.

To the variety of photorectifiers based upon one of these principles a new type of photorectifier has been added. This rectifier consists of a photoconductive CdS powder layer in which the grains are bound together by means of a suitable glass enamel.

The CdS *powder*

The photoconductive CdS powder is prepared by mixing pure CdS powder with equal amounts of Cu and Ga, which act as activator and coactivator respectively. The mixture is fired for some hours in an H_2S atmosphere, at 900°C. This method proved to be more reproducible than firing with a halide flux, and the powder obtained in this way has a high intrinsic photosensitivity. This is seen when pressing this powder into pellets. When the physical defects introduced by the pressing have been removed by a heat treatment, very sensitive elements are obtained, which respond almost linearly both to the applied voltage and to the intensity of the light-excitation. A measure of the sensitivity of the material is the product of mobility μ and the lifetime τ of excess electrons. Here this

* Based on a paper read at the 1960 International Solid State Circuits Conference, Philadelphia, February, 1960.

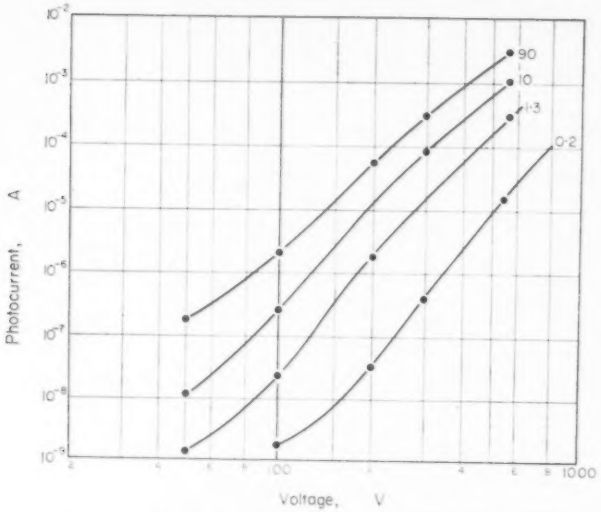


FIG. 1. Plot illustrating the dependence of photocurrent on voltage of a photoconductive CdS layer with a binder for various light intensities (lm/m²).

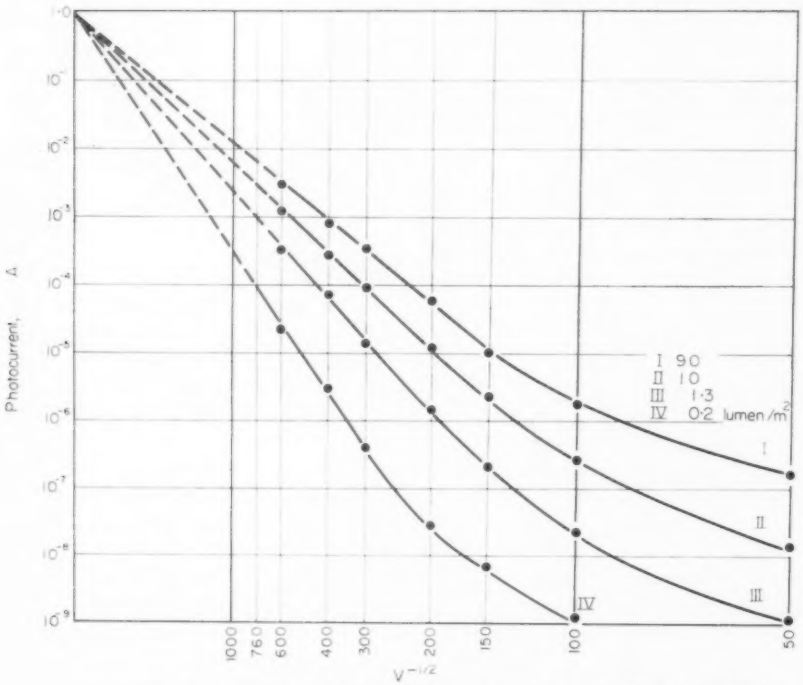
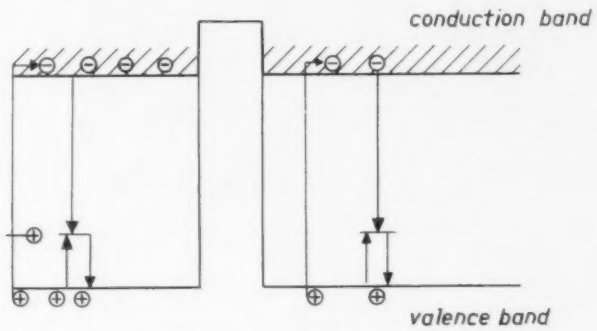


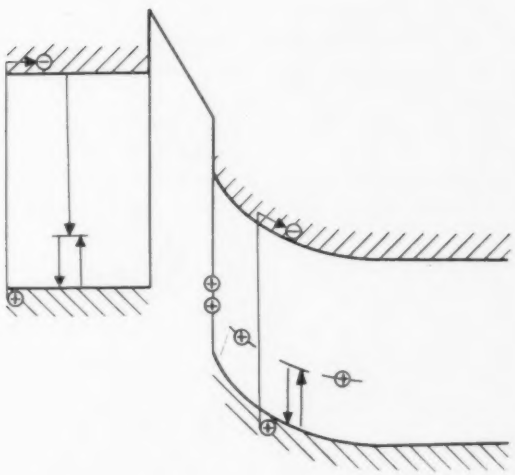
FIG. 2. The same as Fig. 1. Now plotted on a semi-log scale. The abcissa is the inverse square root of the voltage ($V^{-1/2}$).

$\mu\tau$ product is as high as about 1 cm²/V.[1] Large quantities of this type of CdS photoconductor have already been produced.

perties of this type are quite different from those of the pressed and sintered layers. In Fig. 1 the dependence of the photocurrent on the voltage is



(a)



(b)

FIG. 3. Band model for two photoconducting CdS grains with a thin layer of insulator in between. (a) No voltage applied. (b) A d.c. voltage applied.

The preparation and the properties of the photoconductive layer

The photoconductive layer is prepared by mixing the photoconductive CdS powder with a small amount of powdered glass–enamel, spraying the mixture on a substrate, and firing at the melting point of the enamel. The photoconductive pro-

shown for various light intensities. One decade in the voltage corresponds to 4 or 5 decades in the photocurrent. The measurements were performed with a test cell consisting of two Al electrodes, distance 0·5 mm, length 20 mm and with d.c. Owing to the strongly superlinear voltage dependence the sensitivity now is a function of the voltage, but

even near breakdown it is about a factor of 10 lower than that of a pressed and sintered layer. These differences are caused by the fact that the photoconductive properties of the layers with binder are governed by the contact regions between the grains, where the electrons have to pass from one grain to another.

In Fig. 2 the log of the photocurrent vs. the inverse square root of the voltage is plotted for four different light levels. We now find straight lines over 3 to 4 decades which by extrapolations go through the same point at infinitely high voltage. This means that the photocurrent follows a Fowler–Nordheim formula for tunnelling through a barrier in which the field strength is proportional to the square root of the applied voltage.

Calculations of the characteristic of such a barrier were based on the band model, as given in Fig. 3(a).

By excitation, electrons are lifted to the conduction band and nearly all of the holes are trapped in the Cu-levels. Applying a d.c. voltage one gets the situation as given in Fig. 3(b).

An electric field now exists across the barrier and across a part of the positive CdS grain. The barrier is narrowed and tunnelling of the electrons can occur.

With this model it is also possible to explain the well-known experience[2] that such layers are less sensitive to a.c. operation than to d.c., as may be seen from Fig. 4. The light is already on and

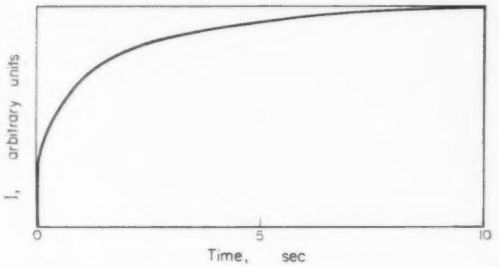


FIG. 4. Build-up of the photocurrent; the voltage was applied at $t = 0$ while the exciting light was already on.

then the voltage is applied at the time $t = 0$. The current rises in a short time to a certain value which is a measure of the a.c. response and afterwards builds up in a rather long time to the larger d.c. value.

The explanation is that, with d.c., a larger space and surface charge can be built up, because with a unidirectional field the holes liberated by the light absorption are swept towards the barrier, which is a rather slow transport process. The result is that a

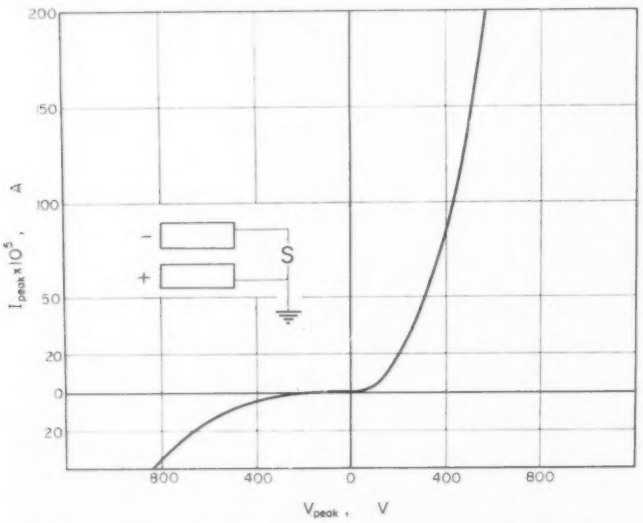
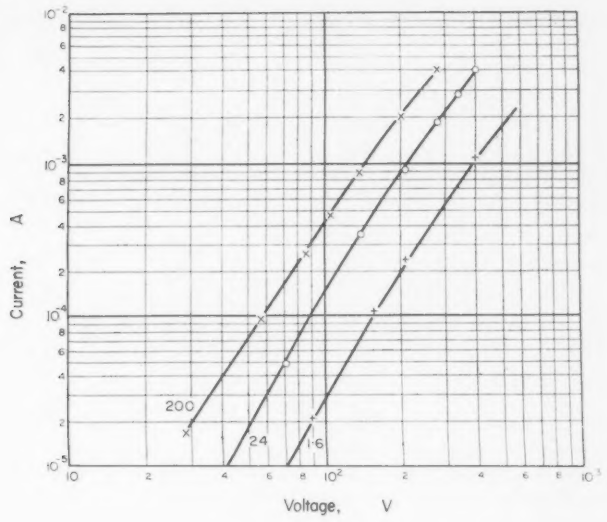
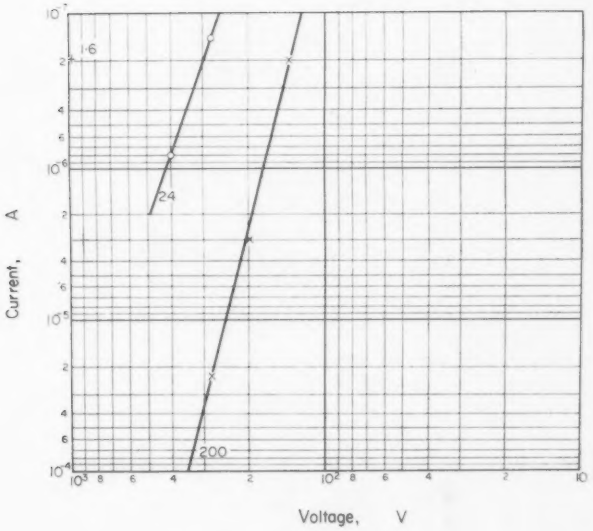


FIG. 5. Dependence of photocurrent on voltage of a formed CdS photorectifier (linear scale); the electrode configuration and the polarity of the voltage during forming are also given.



(a)



(b)

FIG. 6. The same as Fig. 5 on a log–log scale for three different light intensities (lm/m²). (a) Forward direction. (b) Reverse direction.

larger part of the applied voltage appears across the barrier, producing a larger tunnel current. Starting from this explanation of the photoconduction in a photoconductive layer with binder, it is to be expected that the current can be modified by building in a polarization field inside the barrier. This polarization of the glass layer, which forms the barrier, is obtained by a combination of illumination, heating (up to 150–200°C) and passage of a d.c. current through the device. With test cells it is possible to obtain the necessary heating by the passage of the d.c. current itself. When subsequently operated with a.c. at room temperature, rectification occurs with the forward direction opposite to the d.c. polarity during forming. This is shown in Fig. 5.

Both the forward conductance and the rectification ratio depend on the level of illumination. This is shown in Fig. 6(a) and (b). In Fig. 6(a) the dependence of the forward current on the voltage is given for three different light intensities. Thus with a light intensity of 2·8 lm/m² and with a peak voltage of 200 V the forward current is of the order of 1 mA. In Fig. 6(b), the current in the reverse direction is given. This current is much lower than the forward current and its dependence on the voltage is much stronger.

This rectification effect, however, exists with a.c. only. This is shown in Fig. 7. At the time $t = 0$, a d.c. voltage is applied. Just as was the case with a non-formed layer, in the forward direction the current reaches in a very short time a certain value and then builds up more slowly to a higher value. In the reverse direction, however, the current builds up slowly from zero to the same end value. The explanation of these characteristics is as follows. By the combination of heating and application of a d.c. field, the ions in the glass are displaced and a polarization field arises. Before the polarization we therefore assume a symmetrical barrier between the grains, like that in Fig. 3(a), and after the application like that in Fig. 8. The plus and minus signs give the polarity of the d.c. forming voltage. With such an asymmetrical voltage barrier the tunnelling from left to right is decreased, whilst from right to left it is enhanced. It must be borne in mind that the mechanism of the rectification can work only because the positive space charge in the CdS grains can have only a rather low value; thus the charge of the displaced

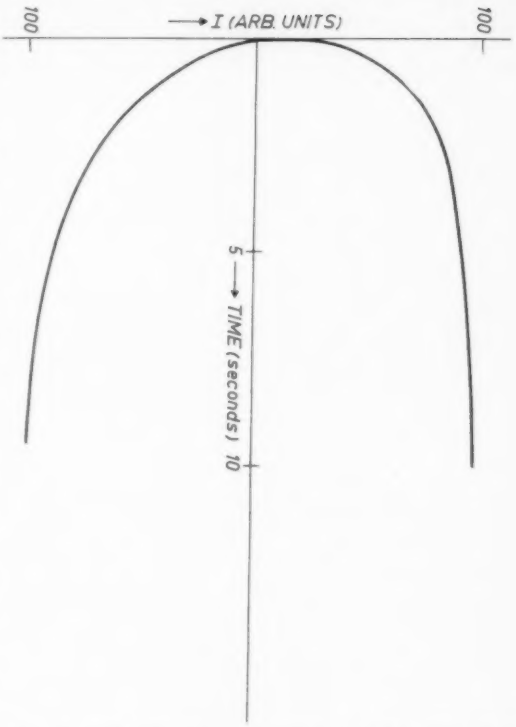


FIG. 7. Build-up of photocurrent with d.c. applied at $t = 0$ after the exciting light was already on. Right-hand scale: forward current. Left-hand scale: inverse current.

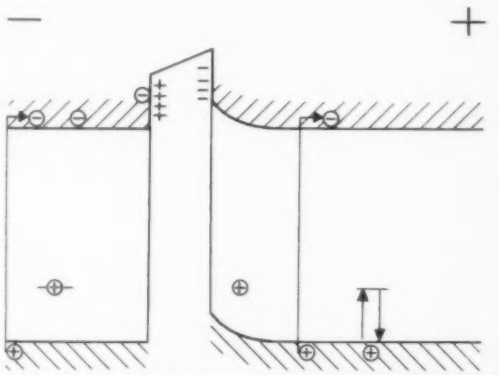


FIG. 8. Band model for the formed photorectifier with d.c. field applied.

ions in the glass is not completely compensated by bulk and surface charges of the adjacent parts of the photoconductor. When the light intensity is in-

creased, the rectification ratio decreases because a larger part of the polarization charge can be compensated. The maximum rectification ratio can be of the order of 10^4.

It is possible to change the direction of the rectification by heating the layer and applying d.c. in the reverse direction. Because the dark conductivity of the layer increases with each heating cycle and the decay time of the photocurrent also increases, this can be done only a limited number of times. The fact that there is no difference between forward and inverse current in the case of d.c. can be explained by assuming that the polarization charge in that case can be completely compensated by surface charges. Another indication that this type of rectification is caused by the polarization of the glass is the fact that the rectification ratio is a function of the composition of the glass. Similar types of glass enamel (all phosphate glasses) but with different cations have been investigated. It was found that at a given forming temperature and current the glasses containing Li showed the highest forward current and the highest rectification ratio and the glasses with K the lowest.

Applications

With this type of photorectifier various arrays have been made. Fig. 9 shows a picture of such an array. The whole system is sprayed on a ceramic support. The CdS layer is grooved as can be seen in Fig. 10; in this way one can make use of surface conduction of the CdS layer. If the layer is not grooved, there remains a high series resistance, since it is difficult to have excitation throughout the bulk. In this array, sets of four conducting lines are used in parallel to diminish the resistance per cross-point. This resistance is nevertheless rather high; with a high rectification ratio, say of 10^4, it can be about 10 kΩ per cross-point. Electrically the response of the CdS is rather fast and thus the array can be used up to a frequency of some megacycles per second, the limit being set by the capacitance of the cross-points. Optically the CdS is much slower; the response time for light varia-

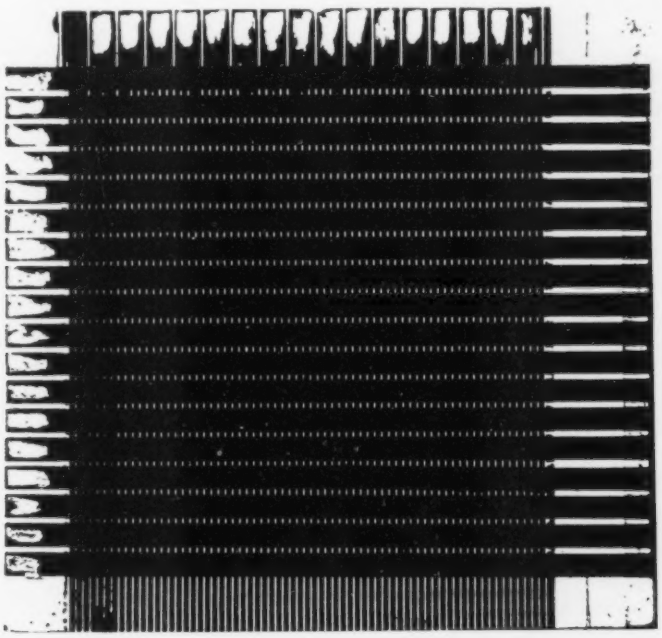


FIG. 9. Picture of a cross-bar photorectifier array.

tions is of the order of 10–1000 msec, depending on the light level. The system can be used for application in such cases where a high voltage from a large number of photorectifiers is needed and the high forward resistance can be tolerated. Such an application may be in matrices for optical reading and switching. The optical input can, for example, be controlled by a punched card placed in front of the photorectifying array. The electrical readout can be performed by applying voltages between chosen horizontal and vertical lines in succession (serial readout). It will be clear that the use of photorectifiers instead of ohmic photoconductors substantially reduces the cross-talk, also when high-ohmic loads are used. When all the cross-points are illuminated except the one that has to be measured (which is the most unfavourable situation) the parasitic current should be smaller by a sufficient amount compared to the signal current flowing when all the cross-points are in the dark except the one measured. It can easily be seen that the maximum useful number of cross-points is equal to the rectification ratio. When they are equal, the parasitic current equals the signal current in the case mentioned and the discrimination disappears.

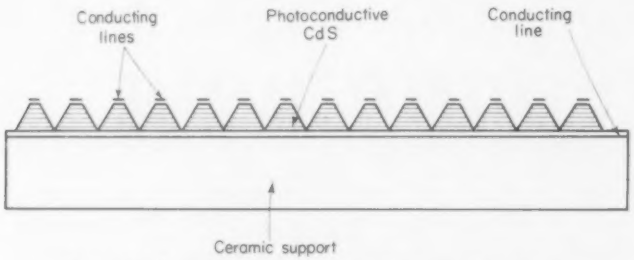


FIG. 10. Cross-section through the array of Fig. 9.

If the voltage source used for scanning has a sufficiently low internal resistance, however, the number of cross-points can be higher and it is only limited by the maximum power to be extracted from that source.

Other applications of the photorectifier matrix may be in combinations with ferro-electric storage matrices and electroluminescent display devices.

REFERENCES

1. N. A. DE GIER, W. VAN GOOL and J. G. VAN SANTEN, *Philips Tech. Rev.* **20**, 277 (1958).
2. B. KAZAN and F. H. NICOLL, *Proc. I.R.E.* **43**, 1888 (1955).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 157–164. Printed in Great Britain

MEASUREMENT OF THE ELECTRICAL RESISTANCE OF METAL–THERMOELECTRIC SEMICONDUCTOR CONTACTS

GWYNN McCONNELL and ROBERT SEHR

The Franklin Institute Laboratories for Research and Development, Philadelphia 3, Pennsylvania

(*Received* 12 *October* 1960)

Abstract—Small contact resistance in thermoelectric junctions relative to the volume resistance is an important factor in the efficiency of the thermoelectric element. A systematic study of different bonding techniques requires a sensitive measuring method. Several methods are discussed and a four-terminal bridge method is described in detail. This method allows one to measure contact resistances of the order of 1 $\mu\Omega$* with an accuracy of ± 5 per cent. Measurements on nickel-plated and unplated soldered junctions of doped Bi_2Te_3 with copper having a contact resistance of 10 $\mu\Omega$ showed no rectification, but a variation of almost 100 per cent in contact resistance around the periphery of the junction.

Résumé—Une faible résistance de contact des jonctions thermoélectriques, relative à la résistance du volume, est un facteur important du rendement d'un élément thermoélectrique. Une étude systématique des différentes techniques de liaisons demande une méthode de mesurage assez précise. Plusieurs méthodes sont présentées et en particulier celle du pont à quatre bornes est décrite en détail. Cette méthode permet de mesurer des résistances de contact d'environ 1 $\mu\Omega$, avec une précision de ± 5 pour cent. Des mesures prises sur des jonctions soudées de Bi_2Te_3 dopé, ayant ou n'ayant pas subi de nickelage, (le cuivre ayant une résistance de contact de 10 $\mu\Omega$) ne démontrent aucun redressement, mais, on note toutefois, un changement d'environ 100 pour cent dans la résistance de contact près de la périphérie de la jonction.

Zusammenfassung—Der Wirkungsgrad eines Thermoelements hängt wesentlich davon ab, dass der Übergangswiderstand der thermoelektrischen Kontakte im Verhältnis zum inneren Widerstand gering ist. Eine systematische Untersuchung verschiedener Kontaktverfahren erfordert eine empfindliche Messmethode. Verschiedene Methoden werden erörtert, und eine Brücke mit 4 Polen wird eingehend beschrieben. Mit diesem Gerät lassen sich Übergangswiderstände von der Ordnung 1 $\mu\Omega$ mit einer Genauigkeit von ± 5 Prozent messen. Messungen an mit Nickel plattierten und unplattierten gelöteten Kontakten von dotiertem Bi_2Te_3 mit Kupfer, die einen Übergangswiderstand von 10 $\mu\Omega$ besassen, zeigten keine Gleichrichterwirkung, doch ergab sich eine Schwankung von fast 100 Prozent an der Peripherie der Kontaktfläche.

† Der Übergangswiderstand wird üblicherweise in Ω-cm²/Fläche ausgedrückt. Da alle unsere Messungen an Proben von 1 cm² Querschnitt ausgeführt wurden, geben wir die Übergangswiderstände in Ohm, wobie zu beachten ist, dass sie sich auf eine Fläche von 1 cm² beziehen.

1. INTRODUCTION

EXCELLENCE in fabricating Peltier junctions has become widely recognized as a critical factor in the design of thermoelectric devices. The efficient application of materials for cooling, heating and power generation is very dependent on the electrical and mechanical quality of metal–thermoelectric semiconductor contacts. It is important to know not only the magnitude of the interface resistance of the bonds, but we must be able also to characterize joining methods as favorable or un-

* Contact resistance is properly expressed in Ω-cm²/area. However, since all our measurements have been made on samples having 1 cm² cross section, we give contact resistances simply in ohms, keeping in mind that these refer to an area of 1 cm².

favorable depending on the exact amount of interface resistance and its reproducibility as well as their physical strength. For a high overall coefficient of performance (COP) of the Peltier junction the junction resistivity r (Ω-cm^2) should be small compared to the volume resistivity $\rho(\Omega$-cm) of the thermoelectric element. The heat current density into the cold junction of a thermoelement of length L and thermal conductivity K, across which exists a temperature difference ΔT, is given by[1]

$$Q = \frac{K}{L}\left[iT - \Delta T - \tfrac{1}{2} - \frac{i^2}{Z}\left(1 + \frac{2r}{\rho L}\right)\right] \qquad (1)$$

where T is the cold-junction temperature, z the figure of merit and $i = SLJ/K$, S being the Seebeck coefficient and J the electric current density. From this equation it is obvious that higher heat-pumping capacity calls for shorter elements. Consequently the bulk resistance of the element will decrease and with it the overall coefficient of performance of the junction unless the interface contact resistance is reduced proportionately.

The detrimental effect of a resistive bond is most serious at the cold junction. Operating currents as high as 50 A will be used and highly concentrated joule heating will subtract from the cooling potential of the junction. The hot end will be less sensitive to the amount of contact resistance, but the power consumed will be increased by poor contact. In a thermoelectric generator, high junction resistances reduce the amount of power which can be delivered to the load.

Bonding thermoelectric junctions by unsophisticated soldering initially yielded contact resistances between 10^{-3} and 10^{-2} Ω. This, of course, is much too high as, for example, the volume resistance of a $\frac{3}{4}$-in. long, $\frac{1}{2}$-in. dia. Bi_2Te_3 element is of the order of 10^{-3} Ω.

2. SURVEY OF LOW-RESISTANCE MEASURING METHODS

As we improved our bonding technique we found it necessary to investigate more sensitive methods to measure low resistances. A brief summary follows:

The principal methods of measuring small contact resistances with high accuracy is the four-probe method in which the sample is connected to two current electrodes in series with a current source and current meter, and two potential probes across the metal–semiconductor interface, connected to a high-impedance voltmeter (Fig. 1). A variation of this method, used by FRITTS[2],

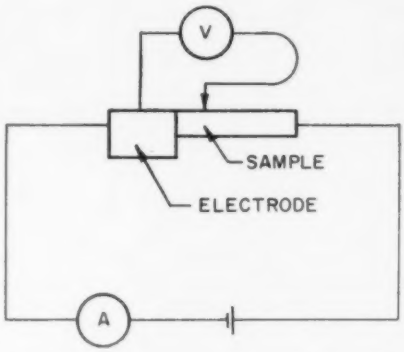


FIG. 1. Four-probe circuit for measuring the junction resistance.

consists of sliding one of the potential probes across the interface and recording the voltage between the potential probes as a function of their distance on an X–Y recorder. The discontinuous voltage jump at the interface divided by the current gives directly the interface resistance.

When the contact resistance encountered is of the order of 1 $\mu\Omega$, a current of about 10 A is required to produce a measurable voltage drop across the potential probes. The inherent inaccuracy of ammeter and voltmeter deflection then limits the precision and therefore the smallest measurable contact resistance.

This limitation can be overcome by the four-probe comparison technique in the Kelvin bridge (Fig. 2). Here the same current passes through the junction resistance and a four-terminal standard resistor while the ratio of the voltage drop across the interface to that across the standard resistor is balanced by an equal ratio across the "ratio arms" of the bridge. Balance is determined with a sensitive null instrument.

All these variations of the four-probe method can be carried out with either d.c. or a.c. However d.c. measurements have the disadvantage to be extremely temperature sensitive due to the thermovoltage arising at the junction from Peltier heating. This effect presents the largest source of error in

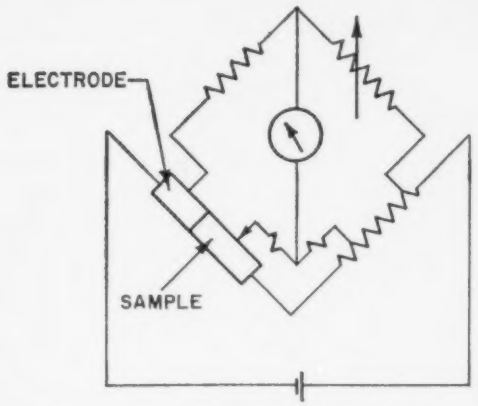


FIG. 2. Four-probe resistance measurement in the Kelvin bridge.

the experiment unless great care is exercised to measure at exactly isothermal conditions.

To demonstrate this, suppose a current of 10 A is flowing through a typical thermoelectric semiconductor whose Seebeck coefficient is about 200 μV/°C. Let the temperature difference between the potential probes be 0·1°C and the junction resistance 10 $\mu\Omega$. The thermal e.m.f. between the potential probes is then $E_s = S\Delta T = 200\times10^{-6}\times0{\cdot}1 = 2\times10^{-5}$ V, while the voltage drop due to the junction resistance $E = IR = 10\times10^{-5} = 10^{-4}$ V. Thus

$$\frac{E_s}{E_R} = \frac{\pm S\Delta T}{IR} = \frac{\pm 2\times10^{-5}}{10^{-4}} = \pm 20\ \% \qquad (2)$$

In other words, a contact resistance of 10 $\mu\Omega$ can be measured only with an accuracy of $\pm 2\,\mu\Omega$ if a small temperature difference of 0·1°C exists across the potential probes. In practice, when measuring in the 10-$\mu\Omega$ range with d.c. the best accuracy has been found to be ± 10 per cent.

3. EXPERIMENTAL

1. *The bridge circuit*

We used a modified Kelvin bridge to measure the metal–semiconductor junction resistance. Fig. 3 shows a schematic diagram of the experimental arrangement. The bridge is powered by a 20 c/s oscillator–amplifier capable of delivering more than 40 W via an impedance-matching balanced transformer. R_2 and R_3 are the adjustable high-

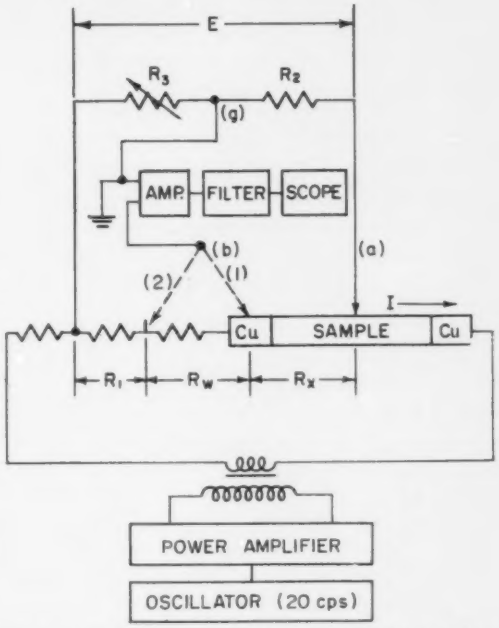


FIG. 3. Schematic diagram of the modified Kelvin bridge used for measuring the junction resistance.

resistance ratio arms, each of the order of $10^3\ \Omega$, grounded at their joining point via the low-input terminal of the detector amplifier. The high-input terminal of the amplifier is alternately connected either to the copper electrode (1) to which the sample is joined, or to the second potential terminal of the four-terminal standard resistor (2). Balancing the bridge in either position yields two equations for the two unknowns, R_x and R_w, from which we can solve for R_x. In position (1) a null is obtained by varying R_3 to R_3' giving

$$\frac{R_3'}{R_1+R_w} = \frac{R_2}{R_x} \qquad (3)$$

while balance in position (2) makes

$$\frac{R_3''}{R_1} = \frac{R_2}{R_w+R_x} \qquad (4)$$

Eliminating R_w gives

$$R_x = \frac{R_1R_2(R_2+R_3'')}{R_3''(R_2+R_3')} \qquad (5)$$

Thus, determining R_x for various distances x and plotting R_x vs. x, a curve is obtained whose intercept with the R_x-axis at $x = 0$ yields the junction resistance.

2. *The sample holder; determination of* $x = 0$

The distance x between the potential probe (a) and probe (b) in position (1) in Fig. 3, is varied by means of a spring-loaded micrometer screw in the sample holder, Fig. 4, which moves the razor-blade probe (a). In order to measure the junction resistance with high precision the zero of the x-axis must be exactly located. This is done in two ways; either optically with the help of a microscope by positioning the razor blade exactly at the visual interface, or, electrically by observing a step-like increase of potential when the razor-blade probe is moved from the copper electrode across the junction. This double check also serves to detect any diffusion of the resistive interface layer in the process of bond aging.

On fresh bonds both methods of determining $x = 0$ are accurate to within 0·02 mm, which in terms of the error resulting from the bulk resistance of the sample is usually less than 1 $\mu\Omega$. However this depends on the test-sample conductivity. The higher that is the smaller will be the fraction of sample resistance to junction resistance. Therefore the slope of the resistance curve intersecting the y-axis will be the less steep the higher the conductivity of the thermoelectric material. This means that the accuracy of determining the junction resistance is proportional to the conductivity of the material. A conductivity of approximately 10^3 $(\Omega\text{-cm})^{-1}$ can cause as much as 4 $\mu\Omega$ error in junction-resistance determination, while good junctions have a resistance of about 10 $\mu\Omega$.

3. *The null detector; importance of high signal-to noise-ratio*

The null detector consists of a Ballantine amplifier (model 300), having an input impedance of 100 kΩ, followed by a two-stage tuned twin-T amplifier, serving as a filter, and an oscilloscope. The latter serves to display the out-of-balance signal of the bridge and to balance the circuit visually. The overall amplification is about 80,000 at 20 c/s; the half width of the pass band is $\frac{1}{2}$ c/s.

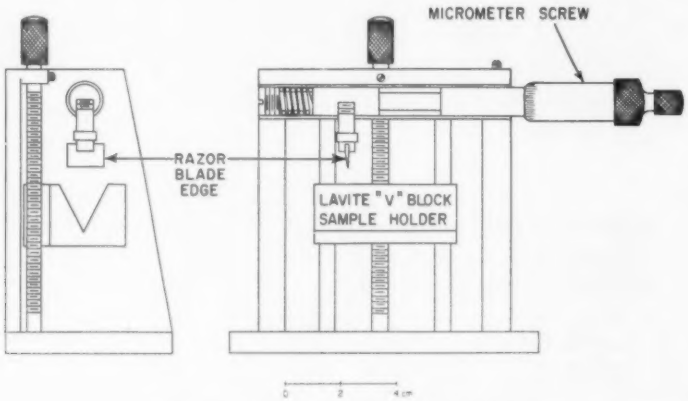


FIG. 4. The sample holder with sliding probe.

As it is shown in the Appendix, the smallest resistance R_x that can be measured with a given relative error $\Delta r = \Delta R_x/R_x$ is

$$R_x = \frac{e}{I \cdot \Delta r - e/(R_1+R_w)} \qquad (6)$$

where e is the smallest detectable voltage at the null-detector input and I the current passing through R_x. From equation (6) it is clear that the minimum measurable resistance R_x decreases with increasing current I. With a given voltage across the bridge (approximately 0·05 V) I is limited by (R_1+R_w) which are in series with the sample. Thus (R_1+R_w) should be made as small as possible. However the presence of the second term in

the denominator in equation (6) requires a relatively large (R_1+R_w) in order to reduce the limit of R_x. With a current $I = 10$ A, a relative error of electrical measurements* of approximately 1 per cent and $e \simeq 10^{-7}$ V, an optimum value $(R_1+R_w) \simeq 10^{-5}\,\Omega$. A calibrated brass bar serves as R_1.

According to equation (6) the lower limit of the measurable junction resistance is proportional to the smallest signal voltage (20 c/s) that the null detector can see. Thus, a very high signal-to-noise ratio is essential to satisfactory operation of the bridge. This was achieved by selective filtering of the tuned detector circuit and by reactance balance in the bridge so that a clean 20 c/s signal emerges. Without this reactance balance† the self-inductance of the high-current-carrying legs of the bridge plus a pick-up from the oscillator leads imposes an out-of-phase signal which cannot be filtered out. Therefore an accurate null balance at very low signal level is impossible.

In order to cancel this out-of-phase voltage an iron-core pick-up coil is inserted in the input lead [(b) in Fig. 3] of the null detector. While varying R_3 this coil is positioned so that it will pick up a signal which is 180° out of phase with the undesirable voltage. As a result of the reactance balance an increase in sensitivity of the bridge of approximately 1000 has been realized so that a resistance variation of $10^{-8}\,\Omega$ can be sensed. The accuracy at the 1-$\mu\Omega$ level is ± 5 per cent.

4. *An alternative technique of determining the junction resistance with the bridge*

Instead of the extrapolation technique described previously, the junction resistance may be determined by measuring the resistance of a section of the semiconductor bar before and after cutting and introducing a bond between the measuring probes. This gives only a resistance value and does not allow one to correct for specimen inhomogeneity.

The material lost during cutting must be allowed for, and, when a copper insert is added, a correction for its length must be made.

We prefer the extrapolation method since the results establish simultaneously the magnitude of two of the principal error contributions. We see first, that any deviation of the plotted points from a straight line is a measure of the inhomogeneity of the specimen. Curvature near the Y-intercept will indicate either non-uniformity of the contact or the effect of random grain orientation.

4. MEASURING

In measuring the contact resistance of thermoelectric junctions we have concerned ourselves mainly with two types of bonds: (1) soldering copper and semiconductor directly, and (2) soldering copper to the electroplated semiconductor. Different solders and fluxes were tried in these experiments. Both have been perfected to about the same degree; both now show contact resistances in the neighborhood of 10 $\mu\Omega$. This may be compared to copper bonds which we have measured and found to be a few tenths of a micro-ohm.

Experiments with ultrasonic soldering produced junctions with resistances as low as 50 $\mu\Omega$. However, this technique was not reproducible. Some bonds were as high as 200 $\mu\Omega$.

One of the largest sources of difficulty in measuring the contact resistances with our method is the electrical inhomogeneity of the sample and a non-uniform contact resistance over the bond area. This effect is illustrated in Fig. 5. Here, measurements were made at 90° intervals on the periphery of a doped Bi_2Te_3 sample. The slope of the resistance–length plot is not quite the same as the sample is rotated, and the straight lines tend to bend near the interface in the first 0·5 mm. These rather critical investigations have shown that the contact resistance may vary as much as 100 per cent around the junction, when the average is about 10 $\mu\Omega$. Fig. 6 also shows this behavior, although to a lesser extent. Here the bond was produced by soldering to an unplated sample of the same material. Also, we note that the contact resistance is not quite proportional to the contact area. Larger samples are slightly more difficult to bond, generally having a higher contact resistance. The contact non-uniformity increases with an increase of area.

* The error of electrical measurement is smaller than the overall error of the junction-resistance measurement. A large contribution to the overall error comes from the interface determination $x = 0$. This can be reduced by doping the sample to high conductivity (see 3.2). Another error source is sample inhomogeneity (see 4).

† This was suggested by Mr. Louis R. Testardi. His participation in setting up the circuit is greatly appreciated.

We have looked for rectification at the junction both with plated and unplated samples. The straightforward d.c. four-probe method was utilized for these experiments with an amplifier across the potential probes. The current–voltage characteristics were displayed on an X–Y recorder. The samples were immersed in a well-stirred isothermal water bath during the measurements. A continually increasing current was passed through the junction in one direction, then in the opposite. In no case did we find a non-linear, i.e. a non-ohmic, junction resistance with currents ranging between 0–50 A.

The bonds were also compared in terms of their mechanical strengths. An Instron tensile testing machine was used in this evaluation. As might be expected, the amount of contact resistance does not show any relationship to the mechanical strength of the bonds. The directly soldered samples showed an average value at failure of 765 lb/in^2 and the plated and soldered samples gave 460 lb/in^2. Some samples ruptured simultaneously with the strongest bonds. These tests show that junctions formed by either technique are probably strong enough for thermoelectric applications.

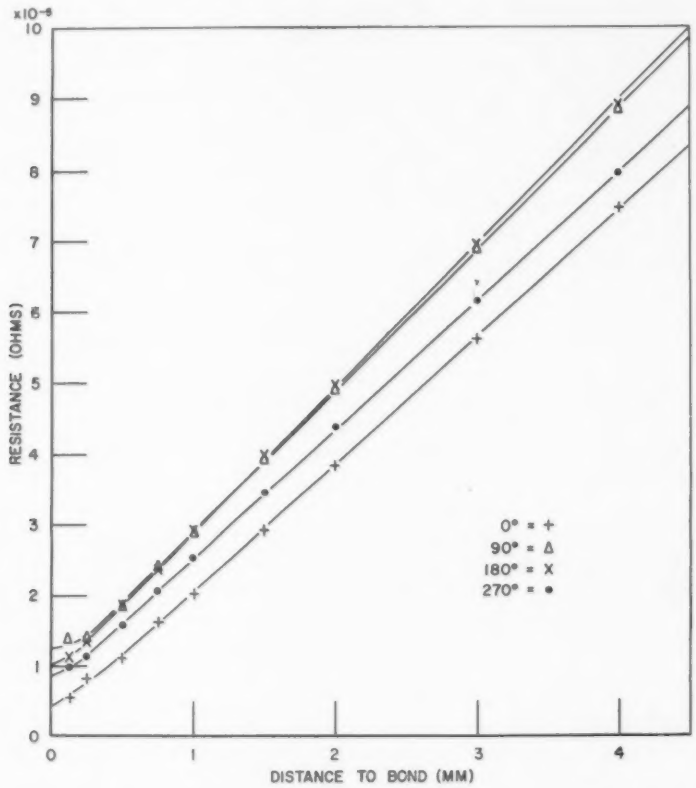


FIG. 5. Resistance profile of a nickel-plated semi-conductor joined to copper. (Contact area 1 cm^2.)

APPENDIX

Referring to Fig. 3, the signal voltage at the terminal (b) of the null detector when in position (1) is

$$E_{b_1} = E\frac{R_x}{R_x+R_1+R_w} \qquad \text{(A1)}$$

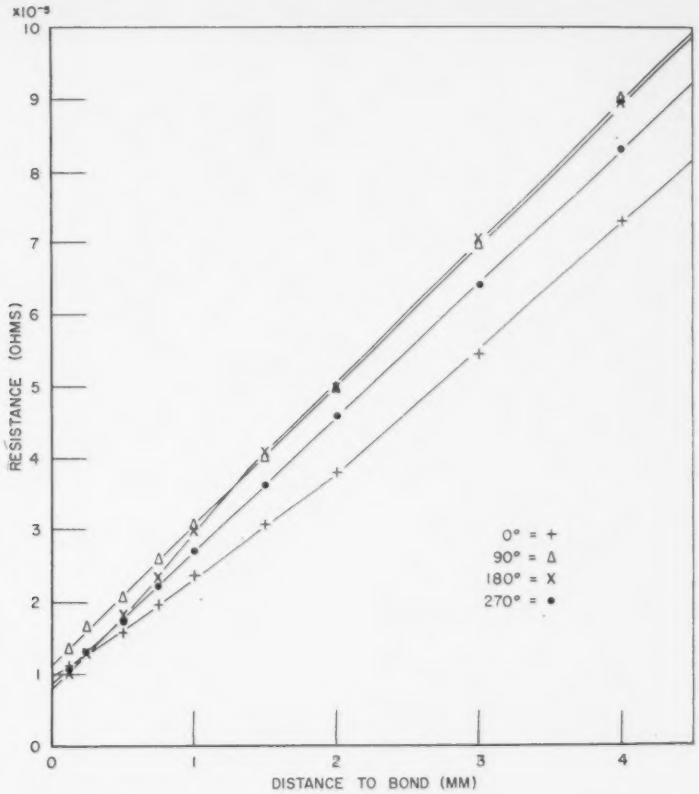


FIG. 6. Resistance profile of an unplated semiconductor joined to copper. (Contact area 1 cm².)

The signal voltage at terminal (g) is

$$E_g = E\frac{R_2}{R_2+R_3} \tag{A2}$$

Thus, the input signal at the null detector is

$$e = E_{b_1}-E_g = E\left(\frac{R_x}{R_x+R_1+R_w}-\frac{R_2}{R_2+R_3}\right) \tag{A3}$$

When the bridge is exactly balanced, $e = 0$ and the bracket expression disappears, i.e.

$$\frac{R_x}{R_x+R_1+R_w} = \frac{R_2}{R_2+R_3} \tag{A4}$$

However, assuming a minimum detectable signal voltage $e \neq 0$, the unknown resistor R_x will be measured with an error $\pm\Delta R_x$. Therefore, instead of (A3) we can write

$$e = E\left(\frac{R_x+\Delta R_x}{R_x+\Delta R_x+R_1+R_w}-\frac{R_x}{R_x+R_1+R_w}\right) \tag{A5}$$

or

$$e = E\frac{R_x}{R_x+R_1+R_w}\left[\frac{1+\Delta R_x/R_x}{1+\Delta R_x/R_x+R_1+R_w}-1\right]$$

Expanding the denominator of the first bracket term in a power series, neglecting second- and higher-order terms and rearranging the right-hand side, we have

$$e = E\frac{\Delta R_x(R_1+R_w)}{(R_x+R_1+R_w)^2} \tag{A6}$$

Introducing the current $I = E(R_x+R_1+R_w)$ through

R_x and the relative error $\Delta r = \Delta R_x/R_x$ we find

$$e = I\frac{\Delta r R_x(R_1+R_w)}{R_x+R_1+R_w} \qquad \text{(A7)}$$

Solving equation (A7) for R_x yields equation (6) which gives the smallest measuring resistance R_x in terms of relative error Δr and network parameters.

REFERENCES

1. F. Donahoe, Private communication (1960).
2. R. N. Fritts, Lead telluride alloys and junctions, in *Thermoelectric Materials and Devices* pp. 143–162. Reinhold, New York (1960).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 165–189. Printed in Great Britain

MECHANISMS OF SPACE-CHARGE-LIMITED CURRENT IN SOLIDS

G. T. WRIGHT

Electrical Engineering Department, University of Birmingham

(*Received* 18 *August* 1960; *in revised form* 17 *October* 1960)

Abstract—The mechanisms of space-charge-limited (SCL) current in solids are discussed. The practical case is taken of a wide band-gap, high-resistivity material containing empty shallow trapping states but in which empty deep trapping states are eliminated by the mechanism of defect compensation described by LONGINI and GREEN (1956). One-dimensional and one-carrier (electron) current through a plane parallel crystal is considered for the case when one contact is ohmic and one contact is blocking.

At small forward voltage, current occurs by the predominant mechanism of carrier diffusion and increases approximately as the exponential of applied voltage; in this range, current is very sensitive to temperature changes. At large forward voltage, current occurs by the predominant mechanism of carrier drift and, after a voltage threshold due to the work-function difference between anode and cathode metals, increases very nearly as the square of applied voltage; this result confirms the simplified analysis of MOTT and GURNEY (1940) and is the solid-state analogue of the three-halves power law for space-charge-limited current in vacuum. In this range current varies as the inverse cube of crystal thickness and is relatively insensitive to temperature changes. Between these two current ranges a smooth transition occurs from a diffusion to a drift mechanism of current and a "virtual cathode" is established in the crystal; there is no evidence for the existence of a negative-resistance region during the transition as predicted by SKINNER (1955). Simple and accurate analytic expressions are derived describing forward current–voltage characteristics in the exponential and square-law ranges; they show that, depending mainly on crystal thicknesses, high forward conductance or high forward resistance can be achieved. With a strongly blocking anode, reverse current is always very small and very high rectification ratios can be achieved.

For current in the square-law range the Fermi-level is nearly constant through the crystal, except near the cathode and anode contacts. This justifies the distinction made by ROSE (1955) between shallow traps, which lie above the Fermi-level and do not affect the form of the current–voltage characteristics, and deep traps, which lie below the Fermi-level and profoundly modify the current–voltage characteristics.

The discussion is illustrated with numerical results calculated on the basis of an electron mobility of 1000 cm^2/V-sec which is intermediate between the value of 200 cm^2/V-sec for cadmium sulphide and 9300 cm^2/V-sec for gallium arsenide. In conclusion, some possible applications are considered for space-charge-limited current in fundamental solid-state research.

Résumé—Les mécanismes du courant à charge d'espace limitée dans les solides sont décrits. Le cas pratique est considéré pour un matérial, ayant une haute résistivité et une large zone interdite, contenant des états de trappe vides et bas, mais dans lequel les états de trappe vides et profonds sont éliminés par le mécanisme de compensation de défaut décrit par LONGINI et GREEN (1956). Le courant d'une dimension et d'un porteur (électron) est décrit pour le cas où un contact est ohmique et un contact est de verrouillage.

A une petite tension de sens avant, le courant est determiné par le mécanisme prédominant de diffusion de porteurs et augmente approximativement en fonction exponentielle de la tension appliquée; dans cette gamme, le courant est très sensible aux variations de la température. A une haute tension de sens avant, le courant est determiné par le mécanisme prédominant d'apport de porteurs et, après une certaine limite de tension, due à la différence de fonction de travail entre les métaux d'anode et de cathode, augmente pratiquement en fonction du carré de la tension appliquée; ce résultat confirme l'analyse simplifiée de MOTT et GURNEY (1940) et est l'analogue à l'état solide de la loi de puissance à 3/2 pour un courant à charge d'espace limitée dans un vide. Dans cette gamme, le courant varie en fonction de l'inverse cubique de l'épaisseur du cristal et est relativement insensible

aux variations de la température. Entre ces deux courants existe une transition régulière d'un mecanisme de diffusion à un mécanisme d'apport de courant et une cathode virtuelle est établie dans le cristal; il n'existe aucune preuve confirmant l'existence d'une région à résistance négative durant la transition comme le prevoyait SKINNER (1955). Des expressions analytiques simples et exactes sont derivées pour decrire les caracteristiques (courant de sens avant–tension) dans les gammes exponentielles et de loi carrée; elles demontrent que, dépendant principalement de l'épaisseur des cristaux, une haute conductance ou une haute résistance du sens avant peut-être obtenue. Avec une anode de fort verrouillage, le courant à sens inverti est toujours très petit et de très hauts rapports de redressement peuvent être obtenus.

Pour le courant dans la gamme de la loi carrée, le niveau Fermi est presque constant à travers le cristal, exception faite près des contacts de l'anode et de la cathode. Ceci justifie la distinction faite par ROSE (1955) entre les trappes basses qui sont situées au-dessus du niveau Fermi et n'affectent pas la forme des caractéristiques courant–tension et les trappes profondes qui sont situées au-dessous du niveau Fermi et qui modifient sensiblement les caractéristiques courant–tension.

L'article contient des résultats numériques calculés sur la base d'une mobilité d'électron de 1000 cm²/V-sec, qui est intermédiare entre la valeur de 200 cm²/V-sec pour le sulfure de cadmium et 9300 cm²/V-sec pour l'arséniure de gallium. Pour conclure, certaines applications possibles dans les recherches fondamentales de l'état solide sont considérées pour un courant à charge d'espace limitée.

Zusammenfassung—Die Entstehung eines durch die Raumladung begrenzten Stromes in Festkörpern wird behandelt. Als praktisches Beispiel dient ein Material mit einer breiten Energielücke und hohem spezifischem Widerstand, das hochliegende leere Einfangsniveaus enthält, bei dem aber die tiefliegenden leeren Einfangsniveaus durch den von LONGINI und GREEN (1956) beschriebenen Mechanismus der Defektkompensation beseitigt wurden. Ein eindimensionaler Einträger-Strom (Elektron) durch einen Kristall mit parallelen Flächen wird für den Fall eines ohmschen und eines Sperrkontakts behandelt.

Bei kleiner Vorwärtsspannung entsteht der Strom in überwiegendem Masse durch Trägerdiffusion und nimmt mit der Spannung nahezu exponentiell zu. In diesem Bereich ist der Strom stark temperaturabhängig. Bei grosser Vorwärtsspannung entsteht der Strom vorwiegend durch Trägerwanderung. Nach Erreichung einer Spannungsschwelle, die durch den Unterschied der Arbeitsfunktionen der Metalle der Anode und Kathode entsteht, wächst der Strom nahezu mit dem Quadrat der aufgewandten Spannung. Dieses Ergebnis bestätigt die vereinfachte Analyse von MOTT und GURNEY (1940) und entspricht für den Festkörper dem Raumladungsgesetz für einen durch die Raumladung begrenzten Strom im Vakuum. In diesem Bereich ist der Strom der dritten Potenz der Kristalldicke umgekehrt proportional und ist gegen Temperaturveränderung relativ empfindlich. Zwischen beiden Bereichen erfolgt ein glatter Übergang von dem auf Diffusion zu dem auf Trägerwanderung beruhenden Strom, und im Kristall entsteht eine "virtuelle Kathode". Das von SKINNER (1955) postulierte Auftreten eines Gebietes mit negativem Widerstand während des Übergangs liess sich nicht bestätigen. Einfache und genaue analytische Beziehungen liessen sich für vorwärtsgerichtete Strom–Spannungs-Kennlinien für die Bereiche mit exponentiellem und quadratischem Verlauf ableiten. Es ergibt sich, dass je nach der Kristalldicke ein hoher Vorwärtsleitwert oder hoher Vorwärtswiderstand erreicht werden kann. Bei stark sperrender Anode ist der Rückwärtsstrom immer klein, und man kann ein hohes Rektifikationsverhältnis erzielen.

Für Ströme im Bereich des quadratischen Verlaufs ist das Ferminiveau im Kristall nahezu konstant, ausser in der Nähe der Kathoden- und Anodenkontakte. Dies bestätigt die Unterscheidung von ROSES (1955) zwischen hochliegenden Einfangsniveaus, die über dem Ferminiveau liegen und die Gestalt der Strom–Spannungs-Kennlinie nicht beeinflussen, und tiefliegenden Einfangsniveaus, die unterhalb des Ferminiveaus liegen und einen starken Einfluss auf die Strom–Spannungs-Kennlinie ausüben.

Der Artikel gibt numerische Ergebnisse, die unter Annahme einer Elektronenbeweglichkeit von 1000 cm²/V-sec berechnet wurden. Dieser Wert liegt zwischen dem von 200 cm²/V-sec für Cadmiumsulfid und 9300 cm²/V-sec für Galliumarsenid. Zum Abschluss wird die Möglichkeit praktischer Anwendungen des durch Raumladung begrenzten Stromes in der Festkörperforschung erörtert.

NOTATION

b $e^2d^2N_0/\epsilon kT$

B $[\beta I_{2/3}(z)+I_{-2/3}(z)]^2/[\beta I_{-1/3}(z)+I_{1/3}(z)]^2$ or $[\beta J_{2/3}(z)+J_{-2/3}(z)]^2/[\beta J_{-1/3}(z)-J_{1/3}(z)]^2$

d crystal thickness

E electric-field intensity

$\mathscr{E}$ reduced electric-field intensity

e electron charge

g^2 integration constant
i current
$I_\nu(z)$ modified Bessel function
$J_\nu(z)$ Bessel function
J current density
j reduced-current density
k Boltzmann's constant
n density of mobile electrons
n_t density of trapped electrons
N_0 effective integrated density of conduction states
N_t density of trapping states
s reduced distance
s_0 integration constant
T temperature
V inner potential difference
$\mathscr{V}$ reduced inner potential difference
V^a applied potential difference
$\mathscr{V}_a$ reduced applied potential difference
W_f Fermi-level
W_t trap depth
y $\exp[-(U-U_0)/2]$
z $\sqrt{(2)}(js+\alpha^2)^{3/2}/3j$ or $\sqrt{(2)}(-js-\alpha^2)^{3/2}/3j$
α^2 integration constant
β integration constant
ϵ permittivity
θ trapping factor
ϕ work function
ψ electron affinity
μ electron mobility

1. INTRODUCTION

THE conduction of electric current in solids normally takes place by the movement of mobile charge carriers which are already present within the material. The situation most usually encountered is ohmic conduction as in metals where the mobile carriers consist of the valence electrons in the top layers of the Fermi sea. These have a high and uniform density n and a constant mobility μ so that under the influence of an applied field E the drift current density is

$$J_{\text{drift}} = \mu enE \qquad (1)$$

In semi-conductors the mobile carrier density is generally much less than in metals and it is possible for significant non-uniform carrier distributions to exist. In this event the partial pressure of the mobile carriers varies from place to place being greatest where the carrier concentration is greatest. Because of these variations in partial pressure, carriers tend to diffuse into regions of low concentration at a rate proportional to the carrier mobility, to the existing concentration gradient of the carriers and to the mean thermal energy of the carriers. Accordingly, the diffusion-current density along the concentration gradient is

$$J_{\text{diff.}} = \mu kT(-dn/dx) \qquad (2)$$

Since both metals and semiconductors contain mobile charge, it is not possible for an excess charge density to build up within the material. Consequently electrical neutrality is maintained throughout the material and the current–voltage relations are relatively simple. However, experiments by ROSE[1] and by SMITH[2] have shown that transient currents can be observed in high-resistivity materials, and ALFREY and COOKE[3] and RUPPEL[4] have observed steady currents which were very small but which were larger than ohmic currents. Recent experiments in the author's laboratories[5–8] have shown that large, steady currents can be obtained in insulating dielectric crystals which, unlike metals and semiconductors, contain no significant density of thermally generated mobile charge carriers. Current is achieved by injecting electrons from an external source into the conduction band of the crystal and applying an electric field to move the electrons through the crystal. An insulating dielectric crystal provided with an electron-injecting contact (cathode) and an electron-collecting contact (anode) has been termed a "dielectric diode"[7,8]; it is the solid-state analogue of the thermionic vacuum diode.

In this paper a theoretical investigation is made of the steady-state mechanisms of space-charge-limited current in solids. The practical case is taken of a thin, plane, parallel crystal provided with an injecting (ohmic) cathode and a blocking anode. It is shown that, at small forward voltages, current is carried predominantly by carrier diffusion and depends on the applied voltage in an approximately exponential fashion; at large forward voltages current is carried predominantly by carrier drift and has an approximately square-law dependence upon applied voltage. High reverse resistances can be obtained and high rectification ratios are possible.

2. BASIC CURRENT EQUATIONS

Since the case of practical interest is the plane, parallel crystal, we shall consider one-dimensional and one-carrier (electron) current. The interior of the crystal is not electrically neutral and the electron density falls from a high value at the cathode to a low value at the anode so that both electric potential gradients and carrier potential gradients exist. Accordingly the total current density is the sum of drift- and diffusion-current densities and is given by*

$$J = \mu enE + \mu kT\frac{dn}{dx} \tag{3}$$

The cathode contact is placed at the origin of x and the anode contact at $x = d$, where d is the crystal thickness. For forward current the electric potential increases positive as x increases; the forward current is thus negative in sign.

In this case of space-charge-limited (SCL) current, the interior of the crystal is not electrically neutral; current is carried by mobile carriers which are injected from an external source and which form an excess space-charge in the crystal. Consequently, Poisson's equation must be satisfied at all points in the crystal and we have

$$\frac{dE}{dx} = -\frac{en}{\epsilon} \tag{4}$$

The current equation (3) and Poisson's equation (4) may be combined conveniently by eliminating the electron density n to give

$$-J = \epsilon\mu E\frac{dE}{dx} + \frac{\epsilon\mu kT}{e}\frac{d^2E}{dx^2} \tag{5}$$

The various constants which appear in this equation may be eliminated now by changing to the dimensionless variables

$$s = \frac{x}{d};\quad \mathscr{V} = \frac{eV}{kT};\quad \mathscr{E} = \frac{edE}{kT} \tag{6}$$

$$j = -\frac{e^2d^3J}{\epsilon\mu k^2T^2};\quad b = \frac{e^2d^2N_0}{\epsilon kT} \tag{7}$$

* The negative algebraic signs of electron mobility μ and electron charge e have been taken into account in deriving this equation; accordingly these quantities are now to be regarded as positive numbers.

In these equations the potential difference V is measured with respect to the cathode metal as zero and it should be noted that the reduced potential $\mathscr{V}$ and the reduced current j are both positive for forward current. Making the appropriate substitutions we obtain

$$\frac{d^2\mathscr{E}}{ds^2} + \frac{1}{2}\frac{d\mathscr{E}^2}{ds} - j = 0 \tag{8}$$

This is the basic equation describing SCL current in solids.

In order to use the current equation (3) and Poisson's equation (4) in this way for describing SCL current in solids, the basic problem is simplified to some extent.

All materials possess small amounts of impurities some of which are able to act as donor or acceptor centres and provide mobile charge carriers in the crystal in addition to those injected as space-charge from an external source. If the impurity content is high as in impurity-activated semiconductors then injected space-charge is only significant in the vicinity of the contacts, and in the bulk of the crystal the mobile carrier density is determined predominantly by the density and depth of the impurity centres. In thin crystals or thin layers or under conditions of high current density, however, injected space-charge can become significant as, for instance, in the p–n junction at high forward bias, but this is a situation not normally encountered. If SCL current is to be observed it is preferable for ohmic conductivity due to carriers generated thermally from impurity or valence levels to be small. This requirement can be expressed quantitatively by saying that the dielectric relaxation time of the crystal, $\rho\epsilon$, must be large compared with t_r, the transit time of the injected carriers through the crystal, in order that the injected charge does not decay significantly while in transit. This condition is implied in the equations given since thermal generation of carriers is neglected. These considerations show that fairly wide band-gap materials are necessary if SCL current is to be observed free from ohmic current. The conditions are barely satisfied, for example, by silicon with a band gap of 1·1 eV, an electron mobility of 1200 cm^2/V-sec and an intrinsic resistivity of about 10^5 Ω-cm, but are satisfied, for example, by gallium arsenide with a band

gap of 1·34 eV, an electron mobility of 9300 $cm^2/V\text{-}sec$* and an intrinsic resistivity of about 10^7 Ω-cm.

Empty trapping levels in the crystal are able to reduce the space-charge current considerably because much of the space charge injected into the crystal is immobilized in traps and is unable to contribute to current; the influence of trapping on SCL current has been discussed by ROSE[1]. If the density of initially empty electron traps is N_t at a depth W_t then, when current occurs, the density of filled traps is

$$n_t = N_t/[1+ \exp\{(W_f - W_t)/kT\}]$$

and the ratio of mobile to trapped electrons is

$$n/n_t = N_0[\exp(-W_f/kT) + \exp(-W_t/kT)]/N_t$$

If the traps lie at least several kT above the Fermi-level then

$$n/n_t = N_0 \exp(-W_t/kT)/N_t$$

such traps may be regarded as shallow, and have the property of leaving free a fixed proportion

$$\theta = n/(n+n_t) = N_0/[N_0 + N_t \exp(W_t/kT)]$$

of the space-charge injected into the crystal. Evidently, such traps will reduce the magnitude of the current but, since the ratio of free to trapped charge always remains constant, will not affect the form of the current–voltage characteristics. Their influence may therefore be taken into account simply by defining an effective permittivity $\epsilon\theta$ instead of ϵ in equation (4); this has the effect of reducing the amount of charge in the crystal which is available for carrying current. Traps which lie at least several kT below the Fermi-level may be regarded as deep traps and are able to alter completely the current–voltage characteristics. For such traps the ratio of free to trapped electrons is

$$n/n_t = N_0 \exp(-W_f/kT)/N_t$$

in this case LAMPERT[9] has shown that the proportion of injected space-charge which remains free is extremely small, varies from place to place in the crystal and depends on the magnitude of current.

Although all materials possess donor centres and deep trapping levels to some extent, the experimental evidence suggests that almost complete mutual compensation of these centres can occur under suitable conditions of crystal growth. This has been discussed by LONGINI and GREENE[10], by KAYALI[11] and by ALLEN[12]. LONGINI and GREENE, in particular, show that if the crystal-growing atmosphere contains impurities capable of producing shallow donor centres in the crystal, these can be incorporated in the crystal lattice much more easily if deep empty electron states exist than if such levels are not present. This is because the deep-lying level provides a state of lower energy for the donor electron than does the donor centre itself, thereby reducing the total energy of donor incorporation by an amount $(W_t - W_d)$ where W_t is the depth of the empty level and W_d is the depth of the donor level. Thus, as the crystal grows, with an unavoidable content of deep electron-trapping states formed by thermal defects and acceptor impurities, the simultaneous incorporation of shallow donor impurities ensures that these deep-lying states become filled. Over-compensation is avoided since, as soon as all the deep electron traps are filled, the total energy required for donor incorporation is increased. LONGINI and GREENE show that, in wide band-gap materials in which $(W_t - W_d)$ can be large if the traps are deep and the donors are shallow, compensation is almost exact. Thus, under suitable conditions of crystal growth, shallow donor centres incorporated in the crystal ionize to become shallow traps, and fill deep traps; the result is that residual ohmic conductivity is much reduced and the density of empty deep traps is reduced to the very low level required for large SCL current to be obtained. Thus, in the practical situation, residual ohmic conductivity is negligible and the influence of deep traps has been eliminated, but shallow traps almost certainly do exist. The equations which have been given and the discussion which follows refer to this practical situation.

It is assumed throughout this discussion that thermodynamic equilibrium is maintained at all times between the electron space-charge and the crystal lattice. At small applied electric field strengths this condition is satisfied; the excess

* This is a calculated lattice mobility (H. EHRENREICH, *Prague Conference on Semiconductors, August*, 1960) but present measured values approach this figure.

energy gained from the field by the electrons between collisions with the lattice is much smaller than the mean electron thermal energy and can be satisfactorily dissipated to the lattice at each collision by excitation of acoustic mode vibrations. This is the situation normally existing in semiconductors, for instance, in which the mobile carrier density is sufficiently large that relatively small applied fields are able to produce large current. Under SCL conditions, however, carrier densities are smaller and applied field strengths are greater; conditions are thus more favourable for the mean carrier temperature to rise above that of the crystal lattice. This is likely to occur when the energy gained from the applied field by the individual carrier in moving a mean free path between collisions, becomes comparable with kT and cannot be wholly dissipated by collisions with the lattice. In germanium at room temperature this situation is reached and the carrier mobility begins to decrease when the applied field is of the order of 10^3 V/cm. At higher fields the carriers are able to dissipate excess energy satisfactorily by exciting optical-mode vibrations of the lattice, and the carrier drift-velocity becomes constant and independent of applied field. However, in the wider band-gap, higher-resistivity materials, such as cadmium sulphide and gallium phosphide which are suitable for the observance of SCL current, carrier mobilities are lower and field dependence of mobility should not occur until rather higher field strengths are reached. The experimental evidence for cadmium sulphide indicates that electron mobility remains constant up to field strengths of at least a few times 10^4 V/cm, so that under normal circumstances it seems justified to accept that the electron space-charge remains in thermal equilibrium with the crystal lattice. However, under SCL conditions, field strengths of this order can be exceeded, particularly under pulse or transient operation, and the possible decrease of carrier mobility at the higher applied voltages should be borne in mind. The influence of field dependence of mobility on SCL current in germanium has been discussed by DACEY[13] and in wide band-gap materials, by LAMPERT[14].

A number of previous studies have been made of electron atmospheres and current in insulators. For instance, VON LAUE[15] and FOWLER[16] have discussed conditions in the crystal in the zero-current case in which the Fermi-level remains horizontal through the crystal. More recent studies have been made by SKINNER *et al.*[17,18], WEST[19], MORANT[20] and particularly SKINNER[21] in connexion with the contact charging of insulators. These authors show that the charge transferred into the insulator and the form of the charge distribution in the insulator are controlled largely by the work functions of the contact metals. They distinguish particularly between the case in which all the transferred charge comes from one contact and the electric potential rises steadily through the insulator and the case in which both contacts contribute to the transferred charge and the electric potential has a maximum within the insulator.

The current-carrying case is more difficult to describe. A formal mathematical solution of the basic equation (8) may be obtained readily in terms of Bessel functions.* However, an evaluation of the two integration constants, which is necessary if a useful physical description of current mechanisms is to be obtained, is not possible in explicit form. A simplified treatment of the current-carrying case has been given by MOTT and GURNEY[23] who have considered the situation appropriate to large applied voltages when carrier diffusion can be neglected. These authors show that under these circumstances the SCL current should follow a square-law dependence on applied voltage; this is the solid-state analogue of the three-halves law for SCL current in vacuum. This approach has the merit of simplicity but does not provide an accurate description of current mechanisms particularly near the cathode and anode contacts and is not applicable in the case of small applied voltages. SHOCKLEY[24] has given an approximate discussion of the movement of mobile carriers over a potential maximum, and SHOCKLEY and PRIM[25] give a more accurate discussion of the same case. These authors simplify the problem by considering current in a crystal of undetermined extent and by assuming that, at large distances, drift mechanisms become predominant so that the approximate solution derived by MOTT and GURNEY is obtained. This approach provides a useful account of physical conditions in the crystal for the case of

* A first integration of equation (8) gives a particular form of Riccati's equation; this is discussed by MCLACHLAN[22].

large currents in which a potential maximum does exist in the crystal. However, it cannot be used to discuss current mechanisms at small or large applied voltages, to discuss physical conditions near the cathode or anode contacts, or to provide an account of current–voltage characteristics.

Current mechanisms in a finite crystal bounded by metallic electrodes have been discussed by SKINNER[26]. This author shows that the current–voltage characteristics are influenced markedly by the nature of the electrodes; in particular, the expected rectification characteristics are obtained if the work function of the cathode and anode used by these authors is complex and does not provide a description of physical conditions in the crystal, of current mechanisms, or a usable account of current–voltage characteristics.

The present paper is an attempt to provide a clear and useful description of SCL current in solids. In particular the two integration constants which arise in the formal mathematical solution of equation (8) are evaluated in a manner which enables the underlying physical processes to be kept in mind. Before discussing the current-carrying case, however, it is appropriate to review briefly the zero-current case, since this enables a

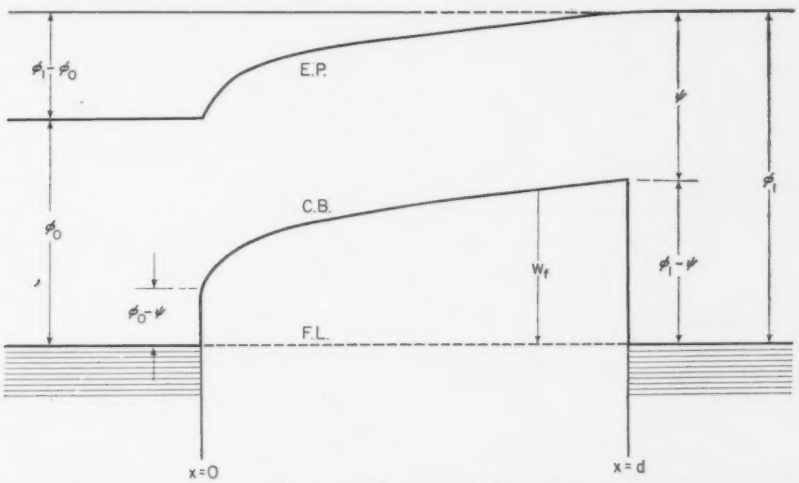


FIG. 1. Energy-level diagram for an insulator crystal bounded by two metallic contacts of unequal work functions. The variations of electric potential, conduction-band level, and Fermi-level are shown respectively by the lines marked E.P., C.B. and F.L. Electric potential is measured positive downwards. ψ = electron affinity of crystal; ϕ_0 = work function of cathode metal; ϕ_1 = work function of anode metal; $\phi_0-\psi$ = cathode potential step eV_0; $\phi_1-\psi$ = anode potential step eV_1; $\phi_1-\phi_0$ = contact potential difference between cathode and anode metals. (W_f = Fermi-level measured downwards from conduction band.)

metals are different. A further and most interesting result is the predicted existence of a negative-resistance region in the current–voltage characteristics. The particular case when cathode and anode metals are identical has been treated by SUITS[27]. In this case no rectification is obtained; the current is proportional to voltage at low applied voltage, but, if space charge is significant, follows a square-law dependence on voltage at high applied voltage. Unfortunately the approach clearer understanding of current mechanisms to be obtained.

3. ZERO-CURRENT CASE

When an insulator which contains no appreciable density of mobile electrons is placed in contact with a metal, electrons diffuse into the insulator until the space-charge field set up in the insulator is sufficient to balance the diffusion of electrons from the metal. The equilibrium situation is

illustrated qualitatively in Fig. 1 which shows the energy-level diagram for an insulator crystal bounded by two metallic contacts. The variation of electric potential through the system is also shown; this is to illustrate in particular that there is no discontinuity of potential at the ideal contact. In this and all subsequent energy-level diagrams the electron energy is measured positive upwards and electric potential is therefore measured positive downwards. The curvature of the conduction band of the crystal near the contact with the metal of lower work function is caused by the space-charge field of the electrons which have diffused from the metal. The Fermi-level in the crystal is measured downwards from the conduction band of the crystal; this is mainly for convenience in the current-carrying case discussed subsequently. For generality we have taken the case in which the contact metals have different work functions. Since the greater charge is transferred into the crystal at the contact with the metal of the lower work function, it is evidently appropriate to take this metal as the cathode.

In equilibrium, the diffusion of electrons away from the contact is balanced by the drift of electrons toward the contact. Accordingly we may put j equal to zero in equation (8) when a first integration gives

$$\frac{d\mathscr{E}}{ds}+\frac{\mathscr{E}^2}{2}=g^2 \tag{9}$$

In this expression g^2 is an integration constant. Under zero-current conditions there is no potential difference between anode and cathode and the Fermi-level is horizontal throughout the insulator crystal. Consequently the density of electrons in the conduction band at any point in the crystal is given by

$$n=N_0\exp(-W_f/kT)=N_0\exp(eV/kT)=N_0e^{\mathscr{V}} \tag{10}$$

Using equations (6) we have

$$\mathscr{E}=-\frac{d\mathscr{V}}{ds};\ \frac{d\mathscr{E}}{ds}=-be^{\mathscr{V}} \tag{11}$$

so that equation (9) gives for the electric-field intensity

$$\mathscr{E}=\pm(2be^{\mathscr{V}}+2g^2)^{1/2} \tag{12}$$

The appropriate sign to be used in this equation depends on the sign of the electric-field intensity. From equation (12) we obtain for the electric potential

$$\exp(\mathscr{V}-\mathscr{V}_0)=\frac{\sinh^2(g^2/2)^{1/2}s_0}{\sinh^2(g^2/2)^{1/2}(s+s_0)} \tag{13}$$

In these equations the quantity eV refers, as usual, to the total energy of the electron. It is not usually necessary or desirable to separate the thermal and electric components of the total energy, but it should be borne in mind that the potential steps eV_0 and eV_1 at the contacts are not necessarily caused by a change of electric potential. This is of relevance when considering boundary conditions. The Fermi-level of the system, in particular the Fermi-level of the cathode metal, is a convenient level from which to measure the electron energy and if this is done the actual electric potential in the crystal relative to the cathode metal is

$$V+(\phi_0-\psi)/e=V-V_0$$

Now the integration constants g^2 and s_0 can be found from the equilibrium boundary conditions on the electron-charge density at the contacts; at cathode and anode we have respectively

$$\mathscr{V}_0=eV_0/kT=-(\phi_0-\psi)/kT$$

and

$$\mathscr{V}_1=eV_1/kT=-(\phi_1-\psi)/kT$$

Unfortunately the constants cannot be expressed explicitly in terms of the boundary conditions and it is therefore necessary to derive approximate expressions.

Consider the case when $\mathscr{V}_0$ is small (ohmic or injecting cathode) and $\mathscr{V}_1$ is large (blocking anode) which is the case to which this paper particularly refers. The electron density in the crystal at the cathode contact is large under these circumstances, but it must evidently decrease rapidly further into the crystal; if this were not so then large amounts of charge could be transferred into the crystal and this is contrary to the observed effects of contact charging of insulators. Thus $d\mathscr{E}/ds$ is significant only near the contact, and consequently the term $2b\exp\mathscr{V}$ in equation (12) is negligible except near the contact. Now $\mathscr{E}\simeq\mathscr{V}_1-\mathscr{V}_0$, so we conclude that the integration constant g^2 is large and positive. Combining now equations (10) and (13), it is evident that s_0 must be small if the electron density is to decrease rapidly away from the contact. Using that g^2 is large and s_0 is small,

equation (13) for the potential may be approximated in the vicinity of cathode and anode respectively and we obtain, to a first approximation,

$$s_0^2 = 2\exp(-\mathscr{V}_0)/b \qquad (14)$$

$$\sqrt{(2)}g = -[\mathscr{V}_1 + \log(b/2\mathscr{V}_1^2)] \qquad (15)$$

It should be remembered that $\mathscr{V}_0$ and $\mathscr{V}_1$ in these equations are negative.

The smallest value which can be used for eV_0 without becoming involved in electron degeneracy is obtained from equation (13) and is shown in Fig. 2. The following typical physical parameters have been used: crystal thickness $d = 10\,\mu$; temperature $T = 300°\text{K}$; $eV_0 = 0{\cdot}1$ eV. Various values of eV_1 have been taken appropriate to the cases of a blocking anode (g^2 positive, $eV_1 = 1{\cdot}0$ eV); a semi-infinite crystal (g^2 zero, $eV_1 = 0{\cdot}472$ eV which is the potential at a distance of $10\,\mu$ into a semi-infinite crystal); and an injecting anode (g^2 negative, $eV_1 = 0{\cdot}1$ eV). The integration constants g^2 and s_0 for the latter two cases can be ob-

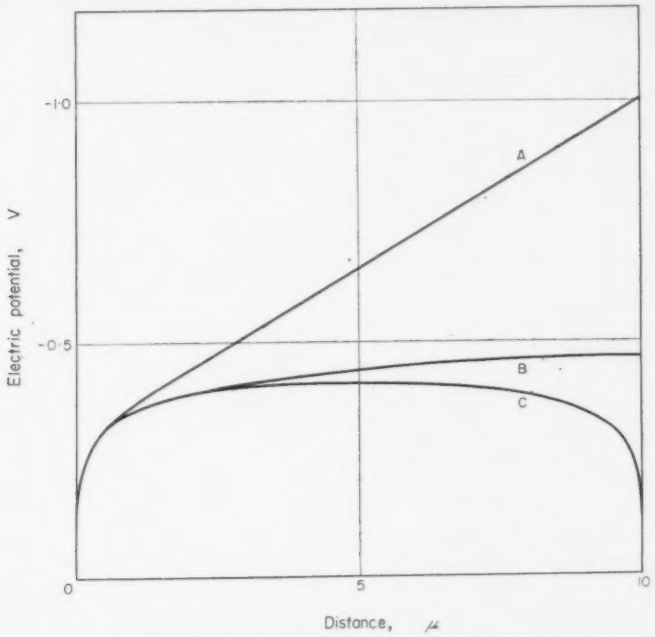


FIG. 2. Variation of conduction band through the crystal for the zero-current case. The zero of reference is the Fermi-level of the cathode metal. Cathode potential step $eV_0 = 0{\cdot}1$ eV; anode potential step $eV_1 = 1{\cdot}0$ eV (curve A), $0{\cdot}472$ eV (curve B) and $0{\cdot}10$ eV (curve C).

at the cathode is about 0·1 eV, and a suitable value of eV_1 to provide a blocking contact is 1·0 eV. It is unlikely of course that eV_0 would be as small as this at a normal contact between a metal and an insulator. However, by suitable formation of the cathode contact it is possible to arrange for the effective value of eV_0 to be small; this is done in the dielectric diode in order that SCL current may be achieved.

The variation of electric potential in the crystal tained by arguments similar to those used for the case of the blocking anode. In this connexion one finds references in the literature[19,21] to the existence of three solutions to the basic zero-current equation. This is unnecessarily confusing; there is only one solution but this transforms from hyperbolic to logarithmic and then trigonometric form according as the integration constant g^2 is positive, zero or negative. However, the diagram shows well that conditions near the cathode are not much

influenced by the nature of the anode. It is also apparent that the injected charge density decreases very rapidly at small distances into the crystal so that most of the injected charge lies within the surface layers of the crystal.

These various forms of electron atmosphere have been discussed in some detail by SKINNER[21] and will not be considered further here since our main interest is in the current-carrying case.

4. CURRENT-CARRYING CASE

The discussion given above of zero-current conditions in the insulator crystal is very useful for visualizing qualitatively the conditions obtaining when current occurs. In this connexion the variation of electric potential and electric field intensity through the crystal are shown qualitatively in Fig. 3 for the particular case of an ohmic cathode contact and a blocking anode contact. For zero current the potential rises steadily and the electric field falls steadily as shown by the full lines. When a positive voltage is applied to the anode, however, electrons are drawn from the cathode space-charge and travel through the crystal towards the anode so that forward current occurs; conditions are now qualitatively as shown by the broken lines. The electron potential energy has a maximum near the cathode which acts as a "virtual cathode" for the emission of electrons into the crystal; this is analogous to the virtual cathode which exists in the thermionic vacuum diode. Up to this point the electron potential energy rises positive like the zero-current case, but beyond this point it falls rapidly, passes through zero and increases negative towards the anode. The electric-field intensity will pass through zero at the position of the virtual cathode and will then increase negative towards the anode.

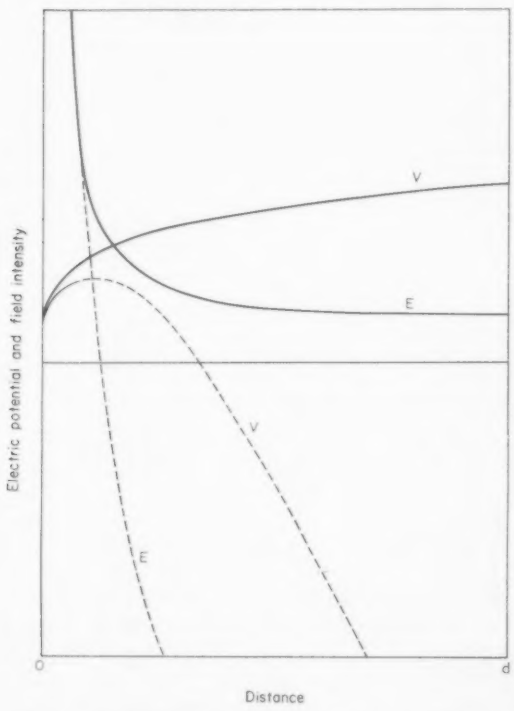


FIG. 3. A qualitative illustration of the variation of electric potential V and field intensity E through an insulator crystal for the zero-current case (full lines) and the current-carrying case (broken lines).

With this diagram in mind we may return to discuss the basic equation (8). The formal mathematical solution will first be given and the appropriate boundary conditions then discussed.

A. *Formal mathematical solution*

The current density is constant throughout the crystal; consequently equation (8) may be integrated directly to give

$$\frac{d\mathscr{E}}{ds} - \frac{\mathscr{E}^2}{2} - (js+\alpha^2) = 0 \qquad (16)$$

The quantity α^2 is a constant of integration. At this stage the algebraic sign and numerical magnitude of α^2 are not known; consequently it is not known whether the quantity $(js+\alpha^2)$ is positive or negative. However, if we now change the variables by writing

$$\log y = \tfrac{1}{2}\int_0^s \mathscr{E}\, ds = -\frac{\mathscr{V}-\mathscr{V}_0}{2} \qquad (17)$$

$$z = \frac{\sqrt{(2)}}{3j}(js+\alpha^2)^{3/2} + i\frac{\sqrt{(2)}}{3j}(-js-\alpha^2)^{3/2} \qquad (18)$$

we may select the real part of z if $(js+\alpha^2)$ is positive and the imaginary part of z if $(js+\alpha^2)$ is negative. The discussion will be continued on the assumption that this quantity is positive and will be transformed as necessary if it should be negative.

Making the substitutions we obtain

$$\frac{d^2y}{dz^2}+\frac{1}{3z}\frac{dy}{dz}-y=0 \tag{19}$$

This equation has the solution

$$y=\text{const.}\times z^{1/3}[\beta I_{-1/3}(z)+I_{1/3}(z)] \tag{20}$$

where β is a second constant of integration. By taking $(js+\alpha^2)$ positive and using the real part of z, the solution has been obtained in terms of Bessel I functions rather than in terms of J functions. This seems more appropriate, since the I functions are non-oscillatory and conditions in the crystal must evidently be non-oscillatory. The constant serves to determine the absolute level of potential and is fixed as soon as the reference zero of potential is decided on. As in the zero-current case, the Fermi-level of the cathode metal will be taken as the zero of electron energy. The electric potential difference

$$(V-V_0)=(kT/e)(\mathscr{V}-\mathscr{V}_0)$$

between any point in the crystal and the cathode metal is then given by

$$\exp(\mathscr{V}-\mathscr{V}_0)=\left(\frac{z_0}{z}\right)^{2/3}\left[\frac{\beta I_{-1/3}(z_0)+I_{1/3}(z_0)}{\beta I_{-1/3}(z)+I_{1/3}(z)}\right]^2 \tag{21}$$

The electric-field intensity in the crystal is given by $\mathscr{E}=2d(\log y)/ds$ from which

$$\mathscr{E}=[2(js+\alpha^2)]^{1/2}\left[\frac{\beta I_{2/3}(z)+I_{-2/3}(z)}{\beta I_{-1/3}(z)+I_{1/3}(z)}\right] \tag{22}$$

The density of electrons in the crystal is given by $bn/N_0=-d\mathscr{E}/ds$ from which

$$\frac{bn}{N_0(js+\alpha^2)}=\left[\frac{\beta I_{2/3}(z)+I_{-2/3}(z)}{\beta I_{-1/3}(z)+I_{1/3}(z)}\right]^2-1 \tag{23}$$

The discussion which follows becomes much less cumbersome if we now define the quantity B by

$$B=\left[\frac{\beta I_{2/3}(z)+I_{-2/3}(z)}{\beta I_{-1/3}(z)+I_{1/3}(z)}\right]^2 \tag{24}$$

so that we may write for the electric-field intensity

$$\mathscr{E}=[2(js+\alpha^2)B]^{1/2}=[2(js+\alpha^2)+2bn/N_0]^{1/2} \tag{25}$$

and for the electron density

$$bn=N_0(js+\alpha^2)(B-1) \tag{26}$$

To proceed further it is necessary to evaluate the integration constants α^2 and β. This can be done through proper choice and utilization of boundary conditions

B. *Boundary conditions*

Under zero-current conditions the electron atmosphere in the crystal is in thermodynamic equilibrium throughout with the electron atmospheres in cathode and anode metals. In particular this is true in the crystal at the actual cathode and anode surfaces. This circumstance was used to provide the boundary conditions for the zero-current case, and the same boundary conditions may be used for the current-carrying case.

Consider first the cathode contact. The cathode metal acts as an effectively infinite reservoir of electrons, and any practical current will not appreciably affect the equilibrium electron atmosphere at the actual cathode contact itself. Thus in the surface of the crystal in contact with the cathode metal, the electron density will always remain constant and equal to $n_0=N_0\exp\mathscr{V}_0$. In the vicinity of the cathode the electron density is large; consequently drift- and diffusion-current tendencies are large. When forward current occurs the diffusion-current tendency becomes larger than the drift-current tendency and we may consider that, near the cathode, forward current is carried by diffusion.

At the anode contact conditions are rather different. When current occurs, electrons move through the crystal towards the anode and evidently the electron density increases considerably over most of the crystal. However, any practical current will not appreciably disturb the equilibrium electron atmosphere in the anode metal, and at the actual anode contact the electron density will remain constant and equal to $n_1=N_0\exp\mathscr{V}_1$. In effect, and because $\mathscr{V}_1$ is large, the anode metal acts as a sink for electrons. Thermal equilibrium will not be reached in the anode metal, of course, over a distance of the order of a collision mean free path, but this is so small compared with any practical crystal thickness that it is quite permissible to assume equilibrium conditions to exist at the actual anode contact. The consequence of this is that in

the crystal near to the anode the electron density falls off very rapidly to the very small value n_1 at the actual anode contact. Near the anode, therefore, there is a large concentration gradient of electrons. Since the electric field cannot change very much over this region, whereas the electron density falls considerably, it is evident that near the anode as well as near the cathode forward current is carried by diffusion.

In the ideal contact, the energy steps eV_0 and eV_1 are due entirely to the discontinuity in the type of crystal lattice which results in a different energy system on either side of the contact. This is the situation, which has been illustrated in Fig. 1, which shows a discontinuity of electron energy at the contacts, although there is no discontinuity of electrostatic potential. In any real contact, however, the situation may be modified by the existence of surface states for electrons and by surface contamination.

Surface states can arise in a variety of ways and it is not possible to do more than make a qualitative estimation of their effects. Tamm surface states will arise at the surface of the free crystal because of the abrupt discontinuity in the periodic potential of the crystal lattice. SHOCKLEY[28] has considered the circumstances under which surface states originate from valence levels and concludes that in this case each surface atom gives rise to one surface state and that these states are half occupied. Discussions of this sort based on the simplified model of a semi-infinite one-dimensional crystal are useful for visualizing qualitatively the situation in a real crystal, and it seems certain that surface states do exist. In semiconductors, for instance, in which the surface charge is compensated by the underlying fixed space-charge of ionized donors or acceptors, BARDEEN[29] has shown that surface states can play a significant part in establishing contact potentials. However, in wide band-gap insulator-type materials in which the Fermi-level lies far from conduction or valence levels, no appreciable curvature of the electron-energy levels near the surface can be caused in this way. Nevertheless, if the crystal surface possessed a high density of empty states, it is presumable that on contact with a metal there could be sufficient transfer of electrons to these to set up a significant space-charge layer of atomic thickness in the crystal surface. The rise of electric potential through this layer would introduce an electrostatic component into the energy step at the contact.

In all real contacts the situation is modified by impurity atoms or molecules adsorbed on to the surface before contact. These surface layers prevent intimate contact between metal and crystal and, moreover, if the ionization potential of the absorbed impurity is lower than the work function of the underlying material so that there is a tendency for electron exchange to occur by ionization of the impurity, or if the adsorbed impurity has a strong electron affinity and so is able to accept an electron from the underlying surface, then an electric dipole layer may be set up at the surface which, as before, would introduce an electrostatic component into the potential-energy step at the contact. A situation similar to this is deliberately produced in the dielectric diode.[6] Here the cathode energy step is made very small by introducing a high density of shallow donor centres into the surface layers of the crystal so that a very thin electric double layer is produced at the cathode contact. This is sufficiently thin as to be transparent to electrons which penetrate the resultant electric potential spike by quantum-mechanical tunnelling.[30] The result is that the conduction band of the crystal at the cathode contact is brought down near to the Fermi-level of the cathode metal; the zero-current inner potential difference across the crystal is now $(\phi_1-\psi-w)/e$, where w is the Fermi-level in the crystal in the highly doped surface layers, instead of $(\phi_1-\phi_0)/e$.

Provided that any potential hump produced in these ways between metal and crystal is very thin, as it will be since adsorbed surface layers are generally only of atomic thickness and electrons in surface states are effectively localized at the surface, the electron atmospheres on each side of the contact will remain in thermal equilibrium although the over-all effective energy step for the contact will not be the true value for the ideal contact. Accordingly, it is quite justifiable to consider that the over-all effect of the contact is to produce an energy step and that the electron atmospheres on either side of the contact remain in thermal equilibrium, although the magnitude eV of the energy step will not necessarily be equal to $(\phi-\psi)$ the true difference between the work function of the metal and the electron affinity of the crystal.

Using these boundary conditions on the electron

density we can in principle obtain the integration constants from equation (23) for the electron density.

In this way we obtain

$$-\beta = \frac{I_{-2/3}(z_0)-(\pm)\sqrt{(B_0)}I_{1/3}(z_0)}{I_{2/3}(z_0)-(\pm)\sqrt{(B_0)}I_{-1/3}(z_0)} \tag{27a}$$

$$-\beta = \frac{I_{-2/3}(z_1)-(\pm)\sqrt{(B_1)}I_{1/3}(z_1)}{I_{2/3}(z_1)-(\pm)\sqrt{(B_1)}I_{-1/3}(z_1)} \tag{27b}$$

where the subscripts 0 and 1 refer respectively to cathode and anode conditions. Either of these equations provides the value of β; by eliminating β between them, the equation defining α^2 is obtained. However, there is an ambiguity in these equations for α^2 and β, since it is not known whether the positive or negative sign is required for $\sqrt{(B)}$. This ambiguity may be resolved by using the fact, from equation (25), that the sign of $\sqrt{(B)}$ is the same as the sign of the electric-field intensity. Consequently the sign required depends on whether cathode or anode conditions are being discussed and on the magnitude and direction of current.

Using these results we may proceed to discuss current mechanisms.

5. CURRENT MECHANISMS IN THE RECTIFYING DIELECTRIC DIODE

For a discussion of the current-carrying case we shall take the practical situation of a crystal provided with an ohmic cathode (n_0 large) and a blocking anode (n_1 small). The forward current case is the more interesting and will be considered first; reverse characteristics will be discussed subsequently.

A. *Forward-current case*

At small applied voltages current is carried predominantly by diffusion and, except at very small currents, follows an approximately exponential dependence upon applied voltage.

(a) *Exponential current range.* When a small forward voltage is applied the current is sufficiently small that conditions in the crystal near the cathode are not very different from the zero-current case. Thus equation (16) should describe conditions which are not very different from zero-current conditions. This shows that for small current $\alpha^2 \simeq g^2$ and, since the maximum value of s is unity, that $j \ll \alpha^2$. Accordingly the quantity $(js+\alpha^2)$ is positive over the whole range of s from cathode to anode. This means that z is positive and real over the whole range of s so that Bessel I functions may be used throughout. Further, since we are considering small currents, z is large and the asymptotic expansions for the I functions may be used throughout. The electric-field intensity remains positive from cathode to anode; consequently the positive sign is required for $\sqrt{(B)}$ in the evaluation of the integration constants.

For small currents, therefore, and to the degree of accuracy required, equation (27) can be expressed

$$-\beta = \frac{\{\sqrt{(B)}-1+[7+5\sqrt{(B)}]/72z\}e^z-[\sqrt{(B)}+1](\cos\pi/6)e^{-z}}{\{\sqrt{(B)}-1+[7+5\sqrt{(B)}]/72z\}e^z+[\sqrt{(B)}+1](\cos\pi/6)e^{-z}} \tag{28}$$

Since z is very large, it is apparent that $-\beta$ is almost equal to unity; it is in fact just less than unity. However, the slight deviation of $-\beta$ from unity is significant, because in equation (21) for the electric potential we are concerned with the very small difference between two Bessel I functions which are extremely large and almost equal in magnitude. Introducing cathode and anode conditions in turn into this equation, equating the two results, and using that $B_1 \simeq 1$, $B_0 \gg 1$ and $z_0, z_1 \gg 1$, we obtain

$$\exp[-2(z_1-z_0)] = \{[\sqrt{(B_1)}-1]/2+1/12z_1\}\times \times[\sqrt{(B_0)}+1]/[\sqrt{(B_0)}-1] \tag{29}$$

This result is valid provided $72z_0 \gg 5\sqrt{(B_0)}$, which is certainly true for small currents. These two equations (28) and (29) may be used to provide the integration constants α^2 and β and hence to obtain a description of physical conditions in the crystal.

The electric potential in the crystal is obtained from equation (21). Using the asymptotic expansions of the Bessel I functions together with the value of β given by equation (28), we obtain

$$\exp(\mathscr{V}-\mathscr{V}_0)=\frac{\sinh^2\{z_1-z_0+\frac{1}{2}\log[(B_1-1)/4+1/12z_1]\}}{\sinh^2\{z_1-z+\frac{1}{2}\log[(B_1-1)/4+1/12z_1]\}} \tag{30}$$

For the particular case of zero current this reduces to equation (13) previously derived for zero-current conditions.

In order to obtain the current–voltage characteristics of the crystal, we put $z=z_1$ in equation (30). Using equation (29) this then gives

$$\exp(\mathscr{V}_a-\mathscr{V}_0)=\frac{B_1-1}{B_0-1}+\frac{1}{3z_1(B_0-1)} \tag{31}$$

In this result $\mathscr{V}_a-\mathscr{V}_0$ is the reduced inner potential difference across the crystal; this is the reduced potential difference between the cathode surface of the crystal and the anode surface of the crystal.

It follows that

$$-J=\sqrt{(2)}\alpha n_1\mu kT[\exp(eV_a'/kT)-1]/d \tag{32}$$

In this equation V_a denotes the external applied voltage between cathode and anode metals (external voltages will be dashed throughout to distinguish them from inner potentials). This result implies the rectifying action of the diode although it is strictly valid only for small forward current. Provided that the quantity α^2 does not vary rapidly, it is evident that the current increases approximately exponentially with the applied voltage. This is a useful result for the thinner crystals and thin insulating layers ($d<1\,\mu$) in which current of practical magnitude can be reached at applied voltages small enough that carrier-diffusion mechanisms remain predominant.

For very small currents, conditions in the crystal will not be very different from zero-current conditions, and we may write $\alpha^2=g^2$. This is consistent with equation (29) which then reduces to the zero-current case provided the current is sufficiently small that $1/12z_1\ll[\sqrt{(B_1)}-1]/2$. This condition requires that $j\ll\sqrt{(2)}gb/N_0$; thus equation (32) can be used with $\alpha^2=g^2$ only if $eV_a'\ll kT$ so that

$$-J=\sqrt{(2)}g\mu en_1V_a'/d \tag{33}$$

For larger but still small currents so that we still have $j\ll\alpha^2$ it is evident, as the current increases, that z_1 becomes smaller and eventually $1/12z_1\ll[\sqrt{(B_1)}-1]/2$. Equation (29) now reduces to

$$j\exp[\sqrt{(2)}\alpha]=4\sqrt{(2)}\alpha \tag{34}$$

This equation provides the value of α^2 to be used in equation (32) when $\sqrt{(2)}gb/N_0\ll j\ll\alpha^2$. By examination of this equation it is found that the largest value of j for which a solution is possible is $j=2{\cdot}68$ corresponding to $\alpha^2=4{\cdot}47$. Using these corresponding values of j and α^2, it is apparent that the requirement $j\ll\alpha^2$ is not satisfied. Accordingly, it is not safe to consider that the exponential current range extends up to values of j greater than unity.

For smaller values of j the solution of equation (34) has two branches, one of which provides decreasing values of α^2 as j decreases while the other provides increasing values of α^2. Considering the decreasing branch first, it is found that it is never possible to have $j\ll\alpha^2$, as required in the small current range. Thus this branch of the solution is not consistent with the original equations and may be disregarded. Considering now the increasing branch it is found that α^2 increases slowly as j decreases and except for values of j near unity, it is true that $j\ll\alpha^2$. This branch may therefore be accepted and we have that the exponential current range lies within the limits $\sqrt{(2)}gb/N_0\ll j<1$.

In the exponential current range discussed here the electric field intensity has remained positive throughout the crystal from cathode to anode. This follows from equation (25), since $j\ll\alpha^2$. Consequently current occurs against the direction of the electric field. Thus the diffusion-current tendency is everywhere greater than the drift-current tendency, and current takes place by diffusion. When the applied voltage is large, however, so that $eV_a'\gg kT$, the drift-current tendency is larger than the diffusion-current tendency. Thus between the small-current and large-current ranges a transition occurs from a predominant diffusion mechanism to a predominant drift mechanism. The current range over which this occurs may be termed the transition range and is discussed below.

(b) *Transition-current range.* As the current rises

through the exponential current range, z_0 and z_1 are decreasing; when j is approaching unity α^2 is quite small and z_0 and z_1 have decreased to quite small values. Accordingly, in the transition-current range it is not possible to use the asymptotic expansions of the Bessel functions. The integration constants α^2 and β cannot be evaluated explicitly and a quantitative account of current mechanisms in this range cannot be given easily. However, it is in this range that the current mechanism changes from being predominantly carrier diffusion to predominantly carrier drift and it is in this range that the negative-resistance characteristics predicted by SKINNER should be found. Thus it is of value to discuss the transition range in some detail in order to obtain an understanding of the physical processes involved.

First we may consider the behaviour of the integration constant α^2. Turning to equation (16) we may write this in the form

$$\mathscr{E}^2 = 2(js+\alpha^2)-2\,d\,\mathscr{E}/ds \qquad (35)$$

Now MOTT and GURNEY[23] have shown that, if the space-charge equations are solved on the basis of a drift-current mechanism alone, $\mathscr{E} = -(2js)^{1/2}$. Electron diffusion does exist, however, and assists forward current. Thus if diffusion mechanisms are taken into account the electric field existing for a given current will be smaller than required on the basis of a drift mechanism alone. Since the quantity $d\mathscr{E}/ds$ is negative, it follows that the quantity α^2 becomes negative in the large-current range and is numerically larger than $d\mathscr{E}/ds$. An estimate of the numerical magnitude of α^2 in the large-current range may thus be obtained from equation (4) which gives $d\mathscr{E}/ds = -bn/N_0$. On the basis of a drift mechanism alone the space-charge capacitance of the diode is $3\epsilon A/2d$ giving that the mean space-charge density of electrons is $3\epsilon V_a'/2ed$ where V_a' is the applied voltage. Combining these estimates we conclude that the average value of $d\mathscr{E}/ds$ is of the order of $3eV_a'/2kT$ showing that, for large currents where the applied voltage is certainly greater than unity, we have $-\alpha^2 > 10^2$. This is only a very rough estimate of the magnitude of α^2 but it is sufficient for our present purpose. We conclude therefore that in the transition-current range the quantity α^2 continues to decrease positive, passes through zero and then becomes large and negative.

For the particular case of $\alpha^2 = 0$ we have $z_0 = 0$ and equation (27a) for β simplifies to give $\beta \simeq 0$. Using this value of β in equation (27b) with $\sqrt{(B_1)} \simeq 1$ we find $z_1 = 1{\cdot}04(2)$. The slight deviation of β from zero and $\sqrt{(B_1)}$ from unity does not affect this result significantly. We thus have $j = 4{\cdot}88$ to provide a first fixed point in the transition current range. Using that $\beta \simeq 0$, $z_0 = 0$, we find from equation (21) that for this value of j the inner potential across the crystal is given by

$$\exp(\mathscr{V}_a - \mathscr{V}_0) = 2\beta^2/[z^{1/3}\Gamma(\tfrac{2}{3})I_{1/3}(z_1)]^2 \qquad (36)$$

When the current is sufficiently large that α^2 has passed through zero and become negative, the imaginary part of z is required [from equation (18)] in the cathode regions of the crystal where s is small enough that numerically $js < \alpha^2$. Under these conditions equation (19) transforms to

$$\frac{d^2y}{dz^2} + \frac{1}{3z}\frac{dy}{dz} + y = 0 \qquad (37)$$

and equations (21–24) for the electric potential, electric-field intensity, electron density and the quantity B, respectively, must be expressed in terms of Bessel J functions instead of I functions.

Since α^2 remains negative for large currents, it is apparent that conditions in the cathode regions of the crystal must be described in terms of the imaginary part of z through the remainder of the transition range and for large currents.

Now we have seen that for large currents the electric field at the anode has passed through zero and become negative. Equation (25) shows that $\mathscr{E}_1$ can become zero only if the quantity B_1 becomes zero which requires that $(j+\alpha^2) = -bn_1/N_0$. Thus while the electric field at the anode is near zero the quantity $(j+\alpha^2)$ is negative and numerically very small; under these circumstances the situation is described in terms of Bessel J functions throughout the whole crystal. Using that $(j+\alpha^2)$ and therefore z_1 are very small while $\mathscr{E}_1$ is near zero, the equations for β can be simplified and the behaviour of this integration constant followed. The detailed examination is lengthy and is therefore omitted, but the interesting result is that β possesses a singularity just before $\mathscr{E}_1$ becomes zero. This is illustrated in Fig. 4, which shows the behaviour of the integration constants α^2 and β through the transition-current range. As the anode

electric field goes to zero, β rapidly increases large and positive and finally becomes positive and infinite; it then changes sign and returns from negative infinity as the anode electric field passes through zero and increases negative. At this singularity of β the electric field at the anode changes sign, and drift mechanisms now begin to assist the forward diffusion current. Accordingly it is convenient to regard this as the point of transition from a predominant-diffusion mechanism of current to a predominant-drift mechanism.

This discussion of the transition range has shown that, for the particular case when $\alpha^2 = 0$, a fixed point can be established at which corresponding values of the integration constants and of the current and applied voltage can be evaluated. This can also be done and a second fixed point established for the particular case when $|\beta| = \infty$ although the detailed derivation of each has been omitted for brevity. However, the potential variation through the crystal for each of these points has been calculated from equation (21) and is illustrated in Fig. 5. The position of the Fermi-level in the crystal has been calculated using equation (23) together with the relation $n = N_0 \exp(-W_f/kT)$ and is also shown in the figure. These curves refer to a crystal 1 μ thick at a temperature of 300°K. A typical value of $\epsilon = 10^{-10}$ F/m has been used for the permittivity of the crystal (dielectric constant $\kappa = 11$) and the electron mobility has been taken as 0·1 m²/V-sec (1000 c.g.s. units) which is intermediate between the value 0·02 m²/V-sec for cadmium sulphide and 0·9 m²/V-sec for gallium arsenide.

In order to make this diagram more useful the potential variation and Fermi-level for zero current and for a current above the transition range are also shown. The values of the integration constants for the latter case have been evaluated as described in the next sub-section. It is interesting to observe from this diagram how physical conditions in the crystal alter as the current mechanism changes from being predominantly carrier diffusion to predominantly carrier drift. In particular, as the current increases from zero through the exponential and transition-current ranges and the electron space-charge spreads into the crystal from the cathode there is very little movement of the

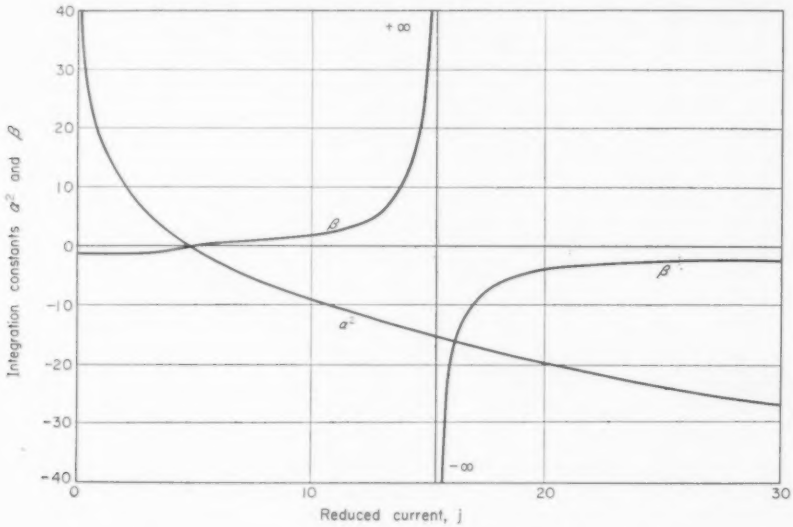


FIG. 4. Dependence of the integration constants α^2 and β upon the magnitude of the reduced-current density j through the transition-current range. The singularity of β marks the transition from a predominant-diffusion mechanism of current towards a predominant-drift mechanism.

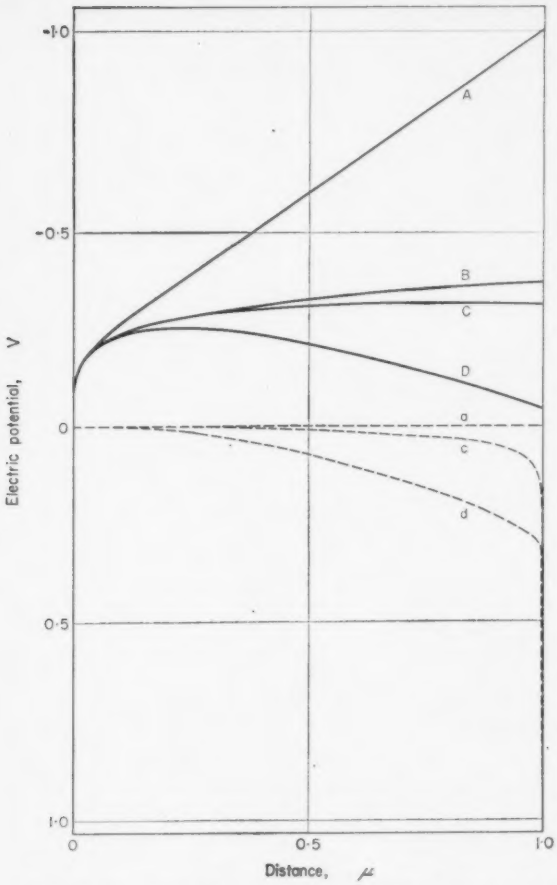


FIG. 5. Variation of conduction band (full lines) and Fermi-level (broken lines) through the crystal for the current-carrying case; crystal thickness 1 μ. Curves A, a: zero-current case; curve B: first transition point, $j = 4{\cdot}78$ giving $-J = 2{\cdot}46$ A/cm^2; curves C, c: second transition point, $j = 15{\cdot}3$ giving $-J = 9{\cdot}16$ A/cm^2; curves D, d: $j = 3\times10^2$ corresponding to $-J = 200$ A/cm^2.

absolute position of the Fermi-level except near to the anode where the electron density falls extremely rapidly. It is not until drift mechanisms become predominant in the large-current range that there is any appreciable movement of the absolute position of the Fermi-level over most of the crystal.

We have seen that for currents above the transition range carrier drift is the predominant mechanism; it is now appropriate to discuss this range in which current follows a very nearly square-law dependence on applied voltage.

(c) *Square-law current range.* We have seen that in this range the quantity α^2 is large and negative but $j \gg \alpha^2$ numerically. Accordingly, the situation must be described in terms of Bessel J functions in the cathode regions of the crystal and in terms of I functions in the anode regions. Since $j \gg \alpha^2$ we have z_1 large, and once again the asymptotic expansions of the I functions may be used in equation (27b) for β to obtain to the degree of accuracy required

$$-\beta = \frac{12z_1e^{z_1} + \cos\frac{\pi}{6}\cdot e^{-z_1}}{12z_1e^{z_1} - \cos\frac{\pi}{6}\cdot e^{-z_1}} \qquad (38)$$

In deriving this expression the negative sign for $\sqrt{(B_1)}$ has been used because $\mathscr{E}_1$ is now negative. This expression is valid provided

$$1/12z_1 \gg [\sqrt{(B_1)}-1]/2 \simeq bn_1/4jN_0$$

which is true in this current range. Since z_1 is large it is evident that $-\beta$ is again almost equal to unity; this time, however, $-\beta$ is slightly greater than unity. As before, the slight deviation of $-\beta$ from unity is significant in evaluating conditions in the anode regions of the crystal, since we are once again concerned with the very small difference between two very large quantities.

Considering now conditions near the cathode, we have, since $j \gg \alpha^2$, that $z \ll z_1$; thus it is sufficiently accurate to take $-\beta$ as equal to unity in the cathode regions of the crystal. Turning now to equation (23) for the electron density it is apparent that the quantity

$$[J_{-1/3}(z_0)+J_{1/3}(z_0)]$$

must be near zero, because B_0 is very large. Since B_0 is so large it is sufficiently accurate to take z_0 to be a zero of this quantity. Taking the first zero to avoid potential infinities in the crystal we find $z_0 = 2{\cdot}38(3)$ giving

$$-\alpha^2 = 2{\cdot}946j^{2/3} \qquad (39)$$

This result provides the value of α^2 in the square-law current range.

Returning to equation (21), the electric potential

in the crystal is given by

$$\exp(\psi - \psi_0) = \left(\frac{z_0}{z}\right)^{2/3}\left[\frac{J_{-1/3}(z_0)+J_{1/3}(z_0)}{\beta I_{-1/3}(z)+I_{1/3}(z)}\right]^2$$

In this expression the quantity

$$[J_{-1/3}(z_0)+J_{1/3}(z_0)]$$

is very small and it is preferable to substitute for it in terms of B_0 from equation (23) for the electron density. We then obtain

$$\exp(\psi - \psi_0) = \left(\frac{z_0}{z}\right)^{2/3}\frac{1}{B_0}\left[\frac{J_{-2/3}(z_0)+J_{2/3}(z_0)}{\beta I_{-1/3}(z)+I_{1/3}(z)}\right]^2 \tag{40}$$

In the vicinity of the cathode where $js < \alpha^2$ numerically, this result must be expressed entirely in terms of Bessel J functions. In the anode regions of the crystal z is sufficiently large that the asymptotic expansions of the I functions may be used. In order to obtain the current–voltage characteristics of the crystal we put $z = z_1$ to obtain

$$\exp(\psi_a - \psi_0) = \frac{2{\cdot}40\pi}{B_0}\left(\frac{z_0}{z_1}\right)^{2/3} z_1 e^{2z_1} \tag{41}$$

In this result $\psi_a - \psi_0$ is, as before, the reduced inner potential across the crystal. It may be noted that this expression is now independent of n_1, the electron density at the anode contact. Thus, in this current range, physical conditions and current mechanisms in the crystal are independent of the precise nature of the anode contact provided that it is a blocking contact, for which the ratio n_1/N_0 is very small. However, it should be remembered that the magnitude of the external applied voltage which is required to produce a given current density does depend on the nature of the anode contact since the external voltage must include the work-function difference between anode and cathode metals.

Combining equations (39) and (41) we obtain

$$-J\left(1-\frac{3}{j^{1/3}}\right)^3 = \frac{9\epsilon\mu}{8d^3}(V_a - V_0)^2\left(1+\frac{kT}{e(V_a-V_0)}\log D\right)^2 \tag{42}$$

where $D = B_0/2{\cdot}40\pi z_1^{1/3} z_0^{2/3}$

This shows that the current follows an approximately square-law dependence upon the applied inner potential difference. The external applied potential difference is given by $V_a' = V_a - V_1$ where $V_1 = -\phi_1/e$. This expression is useful for the thicker crystals ($d > 10\mu$) in which current of practical magnitude is not reached until drift mechanisms become effective; it provides an accurate description of current–voltage relations for large currents for which drift-current mechanisms are predominant. The range of validity of this result is determined by the requirements that $\sqrt{(B_0)} \gg 1$ and $(1+\beta) \simeq 0$; provided that $j > 100$ these requirements are satisfied and equation (42) is accurate. Unfortunately this equation is tedious to use for the calculation of current–voltage characteristics because of the complicated nature of the quantity D. However a considerable simplification can be effected by neglecting the slight dependence of the logarithmic term upon temperature and applied voltage. In this way we obtain

$$-J\left(1-\frac{3}{j^{1/3}}\right)^3 = \frac{9\epsilon\mu V_a^2}{\delta d^3}\left(1+\frac{pkT}{eV_a}\right)^2 \tag{43}$$

This expression is readily manipulated for the purposes of calculation. The reduced-current density j is given in terms of the physical parameters by equation (7). The quantity p is equal to

$$\log\,[d^2e^2N_0/7{\cdot}2\sqrt{(2)}\pi\epsilon kTj^{1/3}(j+\alpha^2)^{1/2}]$$

It varies only slowly with temperature and applied voltage and a value $p = 11$ calculated for $T = 300k$, $j = 1000$, covers practical ranges of temperature and applied voltage. As j increases, the magnitude of B_0 falls and α^2 begins to deviate from the value given by equation (39). However, when the current density, and therefore the applied voltage, is large equation (42) reduces to

$$-J = \frac{9\epsilon\mu}{8d^3}(V_a' + V_1 - V_0)^2 \tag{44}$$

and is independent of α^2. Except for the introduction of the work-function threshold $-(V_1 - V_0)$ this is in fact the approximate result obtained by MOTT and GURNEY on the basis of a drift mechanism alone.

The physical meaning of this result is clear and can be discussed with reference to Fig. 3. In the

vicinity of the ohmic contact to the crystal a space-charge cloud of electrons forms. Current occurs when electrons are drawn from this space-charge reservoir by the applied voltage. Accordingly, electron emission into the crystal occurs from a "virtual cathode" within the crystal. The effective thickness of the crystal is thus reduced by a factor $(1-3/j^{1/3})$. At the virtual cathode the electric potential rises to a maximum (negative) value and then falls rapidly to the anode. Accordingly the effective voltage across the crystal is increased by a factor

$$[1+kT(\log D)/e(V_a-V_0)]$$

Both of these effects are caused by diffusion mechanisms and assist forward current.*

The variation of electric potential through the crystal is shown in Fig. 6 for various increasing values of current through the large-current range. These results have been obtained from equation (40) and show very well the existence of the virtual cathode and the way in which it gradually becomes eliminated as the current increases. These curves refer to a crystal 10 μ thick at a temperature of 300°K for which, as before, the electron mobility $\mu = 0{\cdot}10$ m^2/V-sec and the permittivity $\epsilon = 10^{-10}$ F/m.

The current–voltage characteristic of the crystal is shown in Fig. 7. The abscissa of this diagram gives the external applied voltage and the ordinate gives the square-root of the SCL current density. The diagram shows that there is a small voltage threshold before significant forward current is obtained. This threshold arises basically in the work-function difference $(\phi_1-\phi_0)$ between anode and cathode metals but is reduced to some extent by the effects of carrier diffusion in the crystal. Above this threshold current follows a very nearly square-law dependence upon applied voltage. This is, of course, the solid-state analogue of the three-halves law for SCL current in vacuum. Slight deviations from the exact square-law relation are caused by the contribution of carrier-diffusion mechanisms to the current. At larger currents, for thicker crystals, or at lower temperatures, carrier diffusion becomes less important and the exact square-law characteristic is more nearly approached.

The broken curve in this diagram gives the current–voltage characteristic for a crystal provided with a more highly blocking anode for which $-(\phi_1-\phi_0)/e$ is 2 eV. As expected, the only difference this makes is to shift the curve by 1 V to higher voltages.

The Fermi-level in the crystal may be found by using equation (23) to obtain the electron density in the crystal and then using the relation $n = N_0 \exp(-W_f/kT)$. The variation of Fermi-level is shown in Fig. 8 for the same values of current as used in Fig. 6. The interesting feature of these curves is that the Fermi-level increases rapidly on leaving the cathode then becomes almost constant but increasing slowly over the greater part of the crystal and finally rises extremely rapidly near the anode. This diagram illustrates that near the cathode the electron density falls rapidly from the very large value n_0 to a much lower value which does not vary much over the greater part of the crystal until the vicinity of the anode is reached where the electron density falls extremely rapidly to the very low value n_1. The increase in Fermi-level near the anode occurs extremely sharply over a distance much smaller than an electron mean free path and therefore does not have any physical significance. Because the Fermi-level is so nearly constant over most of the crystal it is justifiable to make a distinction between shallow trapping levels and deep trapping levels as discussed in Section 2; for all practical purposes electron traps at a depth W_t less than 0·3 eV below the conduction band may be regarded as shallow.

In the immediate vicinity of the cathode and anode contacts, current is carried predominantly by carrier diffusion even though carrier drift is the predominant mechanism over most of the crystal. Since physical conditions in the vicinity of the contacts are considerably different from those obtaining over the bulk of the crystal it is of interest to consider the contact regions in more detail.

In the vicinity of the cathode the electron space-charge density is high and the electron atmosphere is very little affected by the occurrence of current. Thus in the cathode region of the crystal, conditions deviate only slightly from the zero-current situation. This may be demonstrated with reference to the electric field at the cathode; for large currents we obtain from equation (35) that

$$\mathscr{E}_0 = (2bn_0/N_0+2\alpha^2)^{1/2} \qquad (45)$$

* This does not imply, of course, that the virtual cathode is at a distance $3d/j^{1/3}$ from the actual cathode or that the height of the potential maximum is $kT(\log D)/e$.

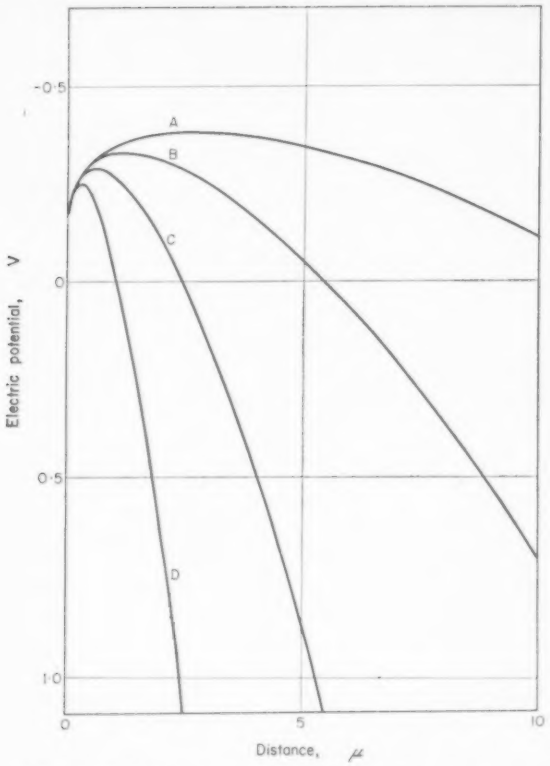


FIG. 6. Variation of conduction band through the crystal for large currents; crystal thickness 10 μ. Curve A: $-J = 0{\cdot}20$ A/cm²; curve B: $-J = 2{\cdot}0$ A/cm²; curve C: $-J = 20$ A/cm²; curve D: $-J = 200$ A/cm².

Now in this current range $\alpha^2 \ll bn_0/N_0$; consequently this is very nearly the same as the zero-current result

$$\mathscr{E}_0 = (2bn_0/N_0 + 2g^2)^{1/2} \qquad (46)$$

The situation near the anode is more interesting. As the current increases the electron space-charge density spreads towards the anode; in the immediate vicinity of the anode, however, the electron density falls rapidly to the relatively very small value n_1. Since z is large near the anode we may use the asymptotic expansions of the Bessel I functions in equations (23) and (27b)* to obtain

* Equation (38) for β is not sufficiently accurate for this purpose; additional terms must be included from the expansion of the Bessel functions.

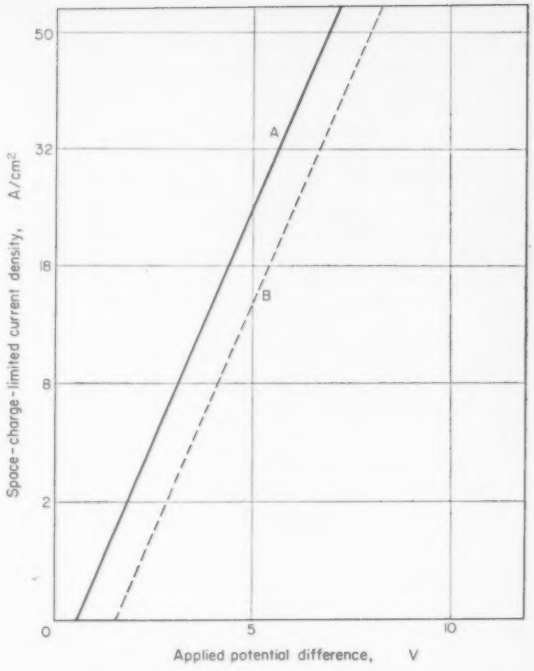


FIG. 7. Forward current–voltage characteristic of space-charge-limited current; crystal thickness 10 μ. Curve A: $(\phi_1 - \phi_0) = 1{\cdot}0$ eV; curve B: $(\phi_1 - \phi_0) = 2{\cdot}0$ eV.

$$\frac{n}{N_0} = \frac{js + \alpha^2}{b}\left(\frac{1}{3z} - \frac{\exp[2(z_1 - z)]}{3z_1}\right) \qquad (47)$$

This expression provides an accurate description of the electron density in the anode regions of the crystal where z is large. At the anode contact itself, however, it gives the value $n_1/N_0 = 0$ instead of the true but very small value $n_1/N_0 = 10^{-16{\cdot}12}$. As we have already seen, however, the precise value of n_1 at the anode contact is immaterial in this current range.

As the current increases through the square-law current range, equation (39) shows that the quantity $-\alpha^2$ increases. The magnitude of B_0, therefore, gradually decreases and the value of z_0 required to satisfy equation (23) gradually decreases. This does not significantly affect the result given by equation (42) because at the larger currents the applied voltage is large. However, this variation of B_0 must be taken into account when discussing current mechanisms in the immediate vicinity of the cathode contact.

If the applied voltage is increased to very large values it will eventually become sufficiently large to move the whole cathode space-charge reservoir across the crystal and the current will saturate and become proportional to applied voltage.

(d) *Saturation-current range.* At very large applied voltages the electric-field intensity in the crystal will be very nearly uniform, and equal to $-V_a'/d$. Turning to equation (25) this can provide a suitable result only if α^2 is positive and very large and if the magnitude of the reduced field intensity is given by $-\sqrt{(2)}\alpha$. We thus have that

$$\mathscr{E} = -\mathscr{V}_a' = -\sqrt{(2)}\,\alpha \qquad (48)$$

in the saturation-current range. In particular, the electric-field intensity at the cathode has decreased from positive values in the large-current range, passed through zero, and is now going large and negative.

Now we have seen that, in the square-law current range, the quantity α^2 increases negative as the current increases. Referring again to equation (25) this shows that the field intensity at the cathode is decreasing and does in fact become zero when $\alpha^2 = bn_0/N_0$. If α^2 became larger than bn_0/N_0 the equation would imply an imaginary field intensity at the cathode. This is not physically permissible and we conclude therefore that the maximum negative value of α^2 is bn_0/N_0 which it reaches as $\mathscr{E}_0$ becomes zero. Subsequently α^2 decreases negative, passes through zero, and increases large and positive into the saturation-current range.

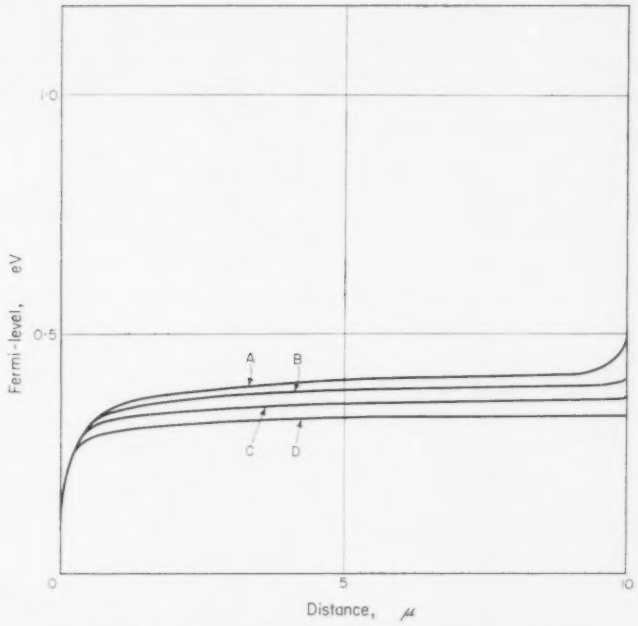


FIG. 8. Variation of Fermi-level through the crystal for currents in the square-law range; crystal thickness 10 μ. Curve A: $-J = 0{\cdot}20$ A/cm^2; curve B: $-J = 2{\cdot}0$ A/cm^2; curve C: $-J = 20$ A/cm^2; curve D: $-J = 200$ A/cm^2.

While the integration constant α^2 is varying in this way z_1 is increasing steadily and equation (38) shows that $-\beta$ remains very nearly equal to unity. Referring now to equation (23) we see, when $\alpha^2 = bn_0/N_0$ and B_0 is zero, that the quantity

$$[J_{2/3}(z_0)+J_{-2/3}(z_0)]$$

is zero. Accepting the first zero as before we have $z_0 = 0{\cdot}685$. This provides a fixed point between

the large-current and saturation-current ranges. Taking again the case of a crystal 10 μ thick at a temperature of 300°K with $\mu = 0{\cdot}1$ m^2/V-sec and $\epsilon = 10^{-10}$ F/m, and using $eV_0 = 0{\cdot}1$ eV, we obtain, when the electric field at the cathode is zero, that $j = 4{\cdot}07\times10^9$ giving $-J = 2{\cdot}74\times10^6$ A/cm^2. Evidently when the crystal is provided with an ohmic cathode the saturation-current range is far above any practical current range; accordingly it will not be discussed further.

So far this discussion has been concerned with forward current; it is now appropriate to discuss the reverse-current case.

B. *Reverse-current case*

With a blocking anode, current in the reverse direction will always be very small. For small reverse voltage in particular, conditions near the cathode will not be significantly different from the zero-current case. Once again therefore we deduce that α^2 is not very different from g^2. At very large reverse voltage, however, the electric-field intensity in the crystal will be effectively uniform and very nearly equal to $-V_a'/d$. Turning to equation (25) for the field intensity this can only provide the correct result if α^2 is very large and positive and if the magnitude of the field intensity is given by $+\sqrt{(2)}\alpha$. At large reverse voltages, therefore, α^2 is very large and positive and we always have $j \ll \alpha^2$. It follows from this that z is everywhere large and real so that Bessel I functions may be used throughout.

Turning to equations (27a) and (27b) for β it is necessary to use the positive sign for $\sqrt{(B_0)}$ and $\sqrt{(B_1)}$ since the electric-field intensity is everywhere positive. However, since the quantities z_0 and z_1 are negative in the reverse-current case, we obtain the same result for β by using the negative sign for $\sqrt{(B)}$ and taking z positive. This case has been evaluated previously to provide equation (38); to the rather greater degree of accuracy required here we have

$$-\beta = \frac{[\sqrt{(B)}+1]e^z - \{\sqrt{(B)}-1-[7+5\sqrt{(B)}]/72z\}\,(\cos\pi/6)e^{-z}}{[\sqrt{(B)}+1]e^z + \{\sqrt{(B)}-1-[7+5\sqrt{(B)}]/72z\}\,(\cos\pi/6)e^{-z}} \qquad (49)$$

In this result of course the quantities $\sqrt{(B)}$ and z are to be regarded as positive because algebraic sign has been taken into account in the derivation.

Using cathode and anode conditions and equating the two expressions for β obtained from equation (49) we have:

$$\exp[2(z_1-z_0)] = [(B_1-1)/4-1/12z_1]\times \times[\sqrt{(B_0)}+1]/[\sqrt{(B_0)}-1] \qquad (50)$$

For the particular case of zero reverse current this expression reduces to the zero-current case discussed in Section 3.

Equations (49) and (50) may be used to provide the integration constants α^2 and β relevant to the reverse-current case. In particular the current–voltage characteristic of the crystal is obtained using equation (21) together with equation (50) to give

$$\exp(\mathscr{V}_a-\mathscr{V}_0) = \frac{B_1-1}{B_0-1} - \frac{1}{3z_1(B_0-1)} \qquad (51)$$

which reduces to

$$-J = \sqrt{(2)}\alpha\mu kTn_1[\exp(eV_a'/kT)-1]/d \qquad (52)$$

For very small reverse current, conditions in the crystal are not very different from zero-current conditions and we may write $\alpha^2 = g^2$. Thus we have the same result as for very small forward current except that the signs of current and applied voltage have reversed. This shows that the current–voltage characteristic has the same slope on either side of zero voltage.

When the reverse voltage is large ($eV_a' \gg kT$) equation (52) shows that $\sqrt{(2)}\alpha = jN_0/bn_1$. Combining equations (50) and (51) we obtain

$$4\exp[-\sqrt{(2)}\alpha] = [\sqrt{(B_0)}+1]^2\exp(\mathscr{V}_a-\mathscr{V}_0) \qquad (53)$$

This equation provides the value of α^2 for large reverse voltage. In particular, when

$$-(\mathscr{V}_a-\mathscr{V}_0) \gg (2bn_0/N_0)^{1/2}$$

we have

$$-\sqrt{(2)}\alpha = (\mathscr{V}_a-\mathscr{V}_0)$$

from which it follows that

$$J = n_1e\mu[-V_a'-(V_1-V_0)]/d \qquad (54)$$

In this expression V_a' is the reverse applied voltage and is negative, and V_1 and V_0 are each negative as usual. This result shows, as expected, that, at large reverse voltages, current is proportional to applied voltage. With a blocking anode, for which the ratio n_1/N_0 is very small, the reverse current is always very small.

6. REMARKS

The experimental achievement of SCL current in cadmium sulphide crystals has shown that the mechanism of defect compensation is very successful in eliminating the influence of empty, deep trapping levels. Possible practical applications for SCL current in the development of high-frequency, low-noise and temperature-insensitive solid-state devices have been discussed elsewhere[6,8,31–33] and it is of interest here to consider some possible applications in fundamental solid-state research.

ROSE[1], LAMPERT[9] and SMITH[34], in particular, have pointed out that in uncompensated crystals the voltage dependence and temperature dependence of SCL current are sensitive to extremely small densities of deep trapping defects. It has been suggested by them that this could be used to provide information about the density and energy distribution of trapping levels. An alternative approach for the investigation of trapping levels, particularly in compensated crystals in which large SCL current can be obtained, may be to study the temperature and frequency dependence of current fluctuations. In this connexion, the intrinsic level of fluctuations caused by variations in stored charge and by variations in transit time should be extremely small under SCL conditions. Accordingly if the trap density is sufficiently high the predominant mechanism of current fluctuations will be variations in the number of mobile carriers due to exchange of carriers between conduction and trapping levels. Further information about trap depths, densities and lifetimes may be obtained by studying the frequency dependence of current, particularly for the case of shallow traps.[31, 33] These methods for studying electron traps are complementary and could form the basis of a useful technique for studying some aspects of the defect solid state particularly in wide band-gap materials.

In compensated crystals the magnitude of current in the square-law range is a direct measure of carrier mobility and can be used to provide the magnitude of this quantity over a wide range of temperature and applied field strength. This is of interest because the density of such additional scattering centres as ionized donor centres can be brought to a minimum and the observed mobility be more truly representative of lattice scattering. In this connexion, the number of carriers available for current is not influenced much by temperature changes; consequently measurements can be made readily over wide temperature ranges. Furthermore, by using the thicker crystals ($d > 10\,\mu$), high electric fields can be applied without resulting in undue power dissipation in the crystal, and the investigation of field dependence of mobility facilitated.

The voltage threshold for forward current in the square-law current range provides a measure of the zero-current inner potential difference across the crystal. In the practical case, in which the cathode energy step is made as small as possible, this gives the magnitude of the anode energy step which for the ideal contact is equal to the difference between the work function of the anode metal and the electron affinity of the crystal. In principle, therefore, this can provide the absolute position of the electron-energy band system of the crystal, information which at present is virtually non-existent for wide band-gap materials and is necessary to complement theoretical studies of electron band structures and to aid understanding of surface and contact phenomena.

Electrical properties such as photoconductivity and breakdown are influenced partially and in some instances predominantly by contact effects. The measurement of forward SCL characteristics can provide information about defects in the crystal and about the magnitude of the anode energy step to assist the interpretation of measurements of photoconductivity, high field conduction, or electric breakdown subsequently made in the reverse direction and for which the anode contact would act as a normal contact.

7. CONCLUSIONS

This discussion of SCL current in solids has produced some clarification of the physics of the metal–insulator contact. Carrier-concentration gradients and space-charge inevitably exist in the vicinity of any contact, and this detailed discussion

of the simplest case, showing that current is carried predominantly by carrier diffusion near ohmic and blocking contacts, provides qualitative information about more complex cases as, for example, the metal–semiconductor contact and the metal–photoconductor contact.

The discussion has shown that a number of current ranges can be defined on the basis of the physical conditions existing in the crystal. Simple and accurate expressions have been obtained describing in particular current mechanisms and characteristics for small forward currents when diffusion mechanisms are predominant, and for large forward currents when drift mechanisms are predominant. The overall current–voltage characteristics are illustrated in Fig. 9 which shows the calculated characteristics of two crystals with thicknesses of $1\,\mu$ and $10\,\mu$, respectively, and having permittivity 10^{-10} F/m and electron mobility $0{\cdot}1$ m^2/V-sec.

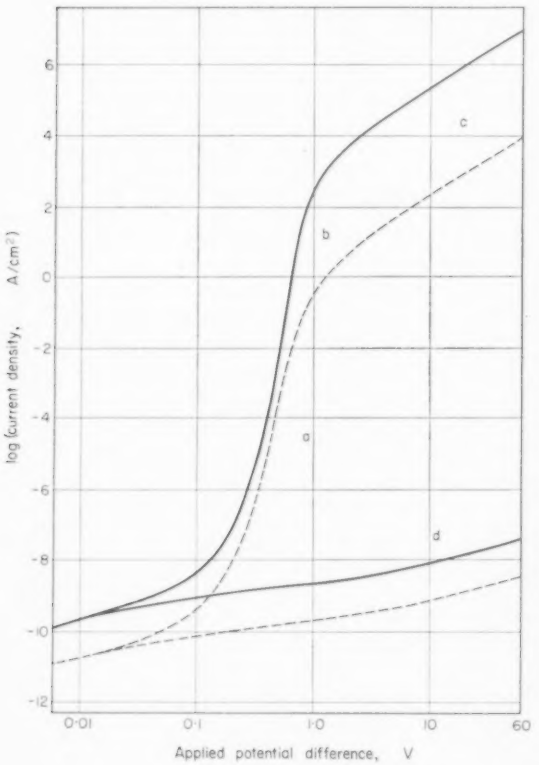


FIG. 9. Current–voltage characteristics of space-charge-limited current; crystal thickness $1\,\mu$ (full lines) and $10\,\mu$ (broken lines). Forward characteristic: (a) exponential range, (b) transition range, (c) square-law range. Reverse characteristic: (d).

At small forward voltage [region (a)] current occurs against the direction of the electric field in the crystal and takes place by the predominant mechanism of carrier diffusion. The current increases approximately exponentially with applied voltage and is very sensitive to temperature. For thin crystals and thin insulating layers ($d < 1\,\mu$), current of practical magnitude can be achieved in this current range.

With larger applied voltages [region (b)] the electric field reverses sign in the anode regions of the crystal and carrier drift now begins to assist carrier diffusion. This is the transition region in which the predominant current mechanism is changing from diffusion to drift and in which the virtual cathode begins to appear. The current increases steadily and smoothly with applied voltage although less rapidly than exponentially and there is no evidence for the existence of a negative-resistance region. The external applied voltage is still less than the work-function difference between anode and cathode metals.

When the applied voltage is of the order, or larger than, the work-function difference between anode and cathode metals [region (c)], the electric field over most of the crystal assists current, and carrier drift is the predominant mechanism of current. In this range, current is of practical magnitude and is very nearly proportional to the square of the applied voltage. Electron emission into the crystal takes place from a "virtual cathode" which exists in the cathode regions of the crystal. Since the effective density of carriers available for current is independent of temperature, current depends on temperature only through the variation of carrier mobility and the extent to which diffusion mechanisms contribute to current. Each of these effects is relatively small, and in any event they act in opposition; consequently in this range current is not very sensitive to temperature. The diagram illustrates well the sensitive dependence of current upon crystal thickness; in this square-law range current is proportional to the inverse cube of crystal thickness.

The Fermi-level in the crystal is nearly constant except near the cathode and near the anode. This

justifies the distinction made by ROSE between shallow traps which lie above the Fermi-level and deep traps which lie below it. This is significant because empty shallow traps merely reduce the magnitude of the current–voltage characteristics without affecting their form, whereas empty deep traps can reduce the magnitude and change the form of the current–voltage characteristics. Near the anode the Fermi-level rises extremely sharply as the electron density falls. However, most of this increase occurs over a distance which is much shorter than an electron mean free path and therefore cannot be used to describe the space-charge electron atmosphere in a real crystal. This means that in the practical case, and providing that the electron density at the anode contact is small compared with the electron density in the central regions of the crystal, the nature of the anode will affect only the magnitude of the forward voltage threshold and will not affect current mechanisms.

At very large forward voltage the whole of the space charge in the crystal is moved across the crystal, and saturation current is obtained proportional to the applied voltage. However, with an ohmic cathode this current range is far above any practical current range.

With a blocking anode for which the ratio n_1/N_0 is small, reverse current is small [region (d)]. As the applied voltage increases the zero-current space charge of electrons in the crystal is gradually swept out through the cathode contact and the resistance of the crystal increases; for large applied voltages it becomes constant and current is proportional to applied voltages. Since the electron space-charge density at the anode is very sensitive to temperature, the reverse current is very sensitive to temperature. Very high reverse resistance and consequently high rectification ratio can be achieved by using anode contacts of high work function.

The numerical calculations used to illustrate this discussion have been made on the basis of an electron mobility of $0{\cdot}1\ m^2/V\text{-sec}$; this is intermediate between the value of $0{\cdot}02\ m^2/V\text{-sec}$ for cadmium sulphide and $0{\cdot}9\ m^2/V\text{-sec}$ for gallium arsenide. Although large and steady SCL current has so far been observed only in cadmium sulphide there seems no fundamental reason why it should not eventually be achieved in a wide range of materials. This would considerably increase the value of SCL current for basic research and for device development.

REFERENCES

1. A. ROSE, *Phys. Rev.* **97,** 1538 (1955).
2. R. W. SMITH, *Phys. Rev.* **97,** 1525 (1955).
3. G. F. ALFREY and I. COOKE, *Proc. Phys. Soc. Lond.* **70B,** 1096 (1957).
4. W. RUPPEL, *Helv. phys. acta* **31,** 311 (1958).
5. G. T. WRIGHT, *Nature, Lond.* **182,** 1296 (1958).
6. G. T. WRIGHT, *Proc. Instn. Elect. Engrs.* **B106,** 915 (1959).
7. A. CONNING, A. A. KAYALI and G. T. WRIGHT, *J. Instn. Elect. Engrs.* **5,** 595 (1960).
8. G. T. WRIGHT, *Electronics* **33,** 59 (1960).
9. M. A. LAMPERT, *Phys. Rev.* **103,** 1648 (1956).
10. R. L. LONGINI and R. F. GREENE, *Phys. Rev.* **102,** 992 (1956).
11. A. A. KAYALI, Thesis, University of Birmingham (1960).
12. J. W. ALLEN, *Nature, Lond.* **187,** 403 (1960).
13. G. C. DACEY, *Phys. Rev.* **90,** 759 (1953).
14. M. A. LAMPERT, *J. Appl. Phys.* **29,** 1082 (1958).
15. M. VON LAUE, *Jb. Radioakt.* **15,** 205, 257, 301 (1928).
16. R. H. FOWLER, *Statistical Mechanics.* Cambridge University Press (1936).
17. S. M. SKINNER, R. L. SAVAGE and J. E. RUTZLER, *J. Appl. Phys.* **24,** 433 (1953).
18. S. M. SKINNER, R. L. SAVAGE and J. E. RUTZLER, *J. Appl. Phys.* **25,** 1055 (1954).
19. D. C. WEST, *J. Appl. Phys.* **25,** 1054 (1954).
20. M. J. MORANT, *J. Appl. Phys.* **25,** 1053 (1954).
21. S. M. SKINNER, *J. Appl. Phys.* **26,** 498 (1955).
22. N. W. MCLACHLAN, *Ordinary Non-linear Differential Equations.* Clarendon Press, Oxford (1950).
23. N. F. MOTT and R. W. GURNEY, *Electronic Processes in Ionic Crystals.* Clarendon Press, Oxford (1940).
24. W. SHOCKLEY, *U.S. Pat.* 6978 (1953).
25. W. SHOCKLEY and R. C. PRIM, *Phys. Rev.* **90,** 753 (1953).
26. S. M. SKINNER, *J. Appl. Phys.* **26,** 509 (1955).
27. G. H. SUITS, *J. Appl. Phys.* **28,** 454 (1957).
28. W. SHOCKLEY, *Phys. Rev.* **56,** 317 (1939).
29. J. BARDEEN, *Phys. Rev.* **71,** 717 (1947).
30. F. A. KROGER, G. DIEMER and H. A. KLASENS, *Phys. Rev.* **103,** 279 (1956).
31. G. T. WRIGHT, *J. Brit. Instn. Radio Engrs.* **20,** 337 (1960).
32. G. T. WRIGHT, *New Scientist* **8,** 1704 (1960).
33. G. T. WRIGHT, *Digest, Solid State Circuits Conference, Philadelphia,* 1961.
34. R. W. SMITH, *R.C.A. Review* **20,** 69 (1959).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 190–193. Printed in Great Britain

NOTES

Contact-etching of semiconductors

(*Received* 1 *December* 1960)

IN THE course of semiconductor processing performed in this laboratory, several etching phenomena were observed which, for want of a better name, will be described as contact-etching. The characteristic feature is the highly increased rate of etching of a semiconductor surface in the immediate vicinity of a metallic or other contact. This effect was first observed with a porous Au chemiplate deposited on a portion of a polished silicon slice. Contrary to expectation, etching in CP-4 resulted initially in preferential dissolution of the region beneath the Au plate and finally in undercutting and removal of the plate. The second observation was the preferential etching of a groove around a Au wire which had been thermocompression-bonded to silicon. Subsequently, a Au rod was found to contact-etch a depression and then a hole in Si. Similarly, a Au wire, when wound tightly around the slice, was observed to contact-etch several grooves in silicon. Fig. 1 schematically shows the arrangement used to drill a hole in a Si chip. A rod holder maintains the Au rod in contact with the semiconductor surface throughout the etching procedure. The force on the rod is small but sufficient to prevent movement of the slice due to agitation of the etchant. The etchant used was CP-4. Fig. 2 shows a cross-section of a hole thus drilled through a slice of Si. The etchant, of course, attacks the entire silicon slice, but preferentially in the vicinity of the Au rod. Fig. 3 shows a photograph of a Si slice with several grooves contact-etched using Au wire. It has also been found that lines may be scribed by moving the Au rod across the surface of a Si slice while immersed in CP-4.

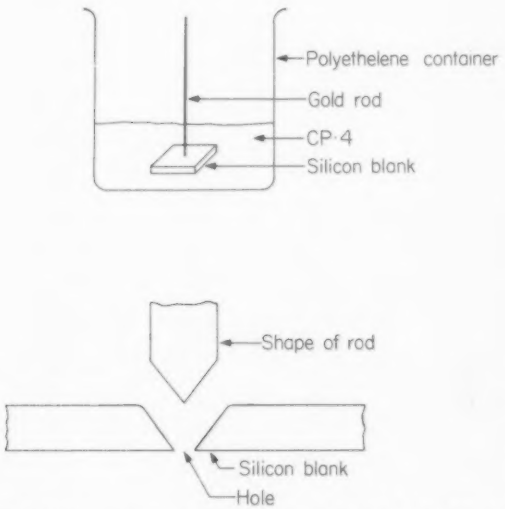


FIG. 1. Arrangement used to drill a hole in a Si blank.

Before enumerating all of the solid-etchant combinations used, a few additional observations will be reported which may offer a clue to the mechanisms involved in contact-etching. The first is the occurrence of contact-etching in a two-metal system. A hole was drilled in a Ag–Ni alloy using Au as the contact material and HNO_3 as the etchant. Secondly, contact-etching was observed using a pointed lucite rod on Si with CP-4 as the etchant. While most of the experiments refer to heterogeneous systems, one observation was made using a Ge–Ge system. In this case one piece of Ge was formed into a point which was then used to contact-etch a depression in a slice cut from the same Ge crystal.

If a conical contact cross-section is used, a replica of the contact shape is obtained. However, sharp edges cannot be maintained at the surface. Best results have been obtained with contact metals which are not attacked by the etchant, although other materials have also been used. Several semiconductor materials were used in this work, encompassing both *n*- and *p*-type and low and high resistivity. These effects can undoubtedly be employed with other semiconductors and with different contact materials and etchants. Table 1 lists the semiconductors, the contact materials and the etchants employed.

The observations reported here have essentially

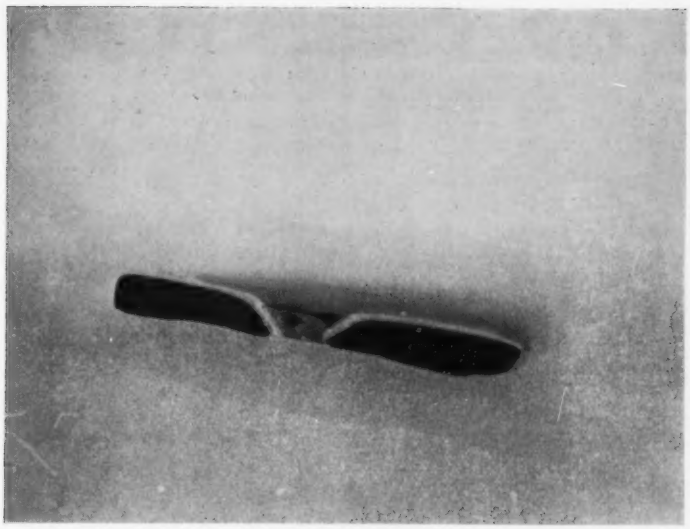


FIG. 2. Photograph of the cross-section of a hole drilled in Si.

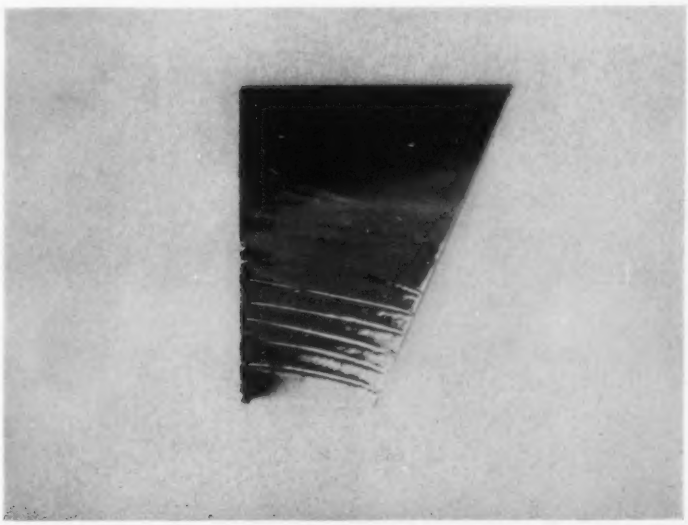


FIG. 3. Photograph of a Si slice with several grooves contact-etched.

[*facing p.* 190

Vol. 2
1961

Table 1. *Materials and etchants used in contact-etching*

Substrate	Contact material	Etchant
Ge*	Au	CP-4
Ge*	Ag	CP-4
Ge*	Ni	CP-4
Ge*	Steel	CP-4
Ge*	Lucite	CP-4
Ge*	Ge	CP-4
Si†	Au	CP-4
Si†	Ag	CP-4
Si†	Ni	CP-4
Si†	Steel	CP-4
Si†	Lucite	CP-4
InSb	Au	dilute CP-4
GaAs	Au	dilute CP-4
Ag–Ni	Au	HNO_3

* *n*-type 15 Ω-cm.

† *n*-type 1·5 Ω-cm, 300 Ω-cm; *p*-type 13 Ω-cm, 1000 Ω-cm, 5000 Ω-cm.

one characteristic in common, namely, preferential etching in the presence of a second contacting material. This does not necessarily imply that one mechanism is responsible for all of the effects observed. For example, the formation of local electrolytic cell action is suggested by the observations with the several metallic contacts on Si and Ge. Cell action is less probable with Ge–Ge and with the lucite–Si system. The effect of pressure must be considered when point-contact elements are used. The possibility of a rod making contact with a small area thus exerting large pressure is also possible. However, the large-area Au plate suggests that pressure is not required. While catalytic action may not be ruled out, the variety of metals used with Ge and Si (and the absence of negative effects) does not suggest such behavior. The presence of the contact may alter the etching conditions as, for example, the temperature, the diffusion and, particularly, the concentration gradients of the reaction products. The extent of the present work does not permit any definite conclusions concerning the mechanisms involved nor has any effort been made to optimize the etching conditions.

Contact-etching does offer an attractively simple technique for fabricating thin base regions for transistors as is done with jet-etching.[1] While no attempt has been made to control the penetration and hence the final thickness, mechanical or optical techniques can undoubtedly be devised. Contacts with circular cross-sections may be replaced with other cylindrical geometries to permit machining of complex shapes into semiconductor bodies. Drilling of small holes through semiconductor chips together with such techniques as diffusion may permit the realization of devices such as the analog transistor.[2]

Acknowledgements—The writers have profited greatly from discussions with Professor J. W. Irvine, Jr., of M.I.T., Mrs. M. C. Lavine of Lincoln Laboratory, and R. C. Ellis, Jr., of Raytheon. They also acknowledge the assistance of D. Davis in carrying out this work.

Research Division W. Rindner
Raytheon Company J. M. Lavine
Waltham, Mass.

References

1. J. W. Tiley and R. A. Williams, *Proc. I.R.E.* **41,** 1706 (1953).
2. W. Shockley, *Proc. I.R.E.* **40,** 1365 (1952).

Control of Lifetime and Diffusion Depth in Semiconductor Devices

(*Received* 27 *December* 1960; *in revised form* 23 *January* 1961)

In the manufacture of diffused semiconductor devices which involve minority-carrier flow, such as transistors or photovoltaic cells, it is necessary that the lifetime of the material be not too much degraded. The degradation appears usually to be due to the accidental introduction of heavy metal impurities such as copper in germanium or gold in silicon. Two methods are used to prevent it: the presence during the diffusion of a second phase in which the impurity is preferentially soluble,[1] and the subjection of the material to a slow cooling process during which the lifetime killers are precipitated and cease to be electrically active.[2] The latter process involves further diffusion during the anneal which at present contributes to the uncertainty in diffusion depth. This Note describes how a specific correction may

be made for diffusion during the anneal. Other sources of uncertainty, not further considered here, include the masking effect of surface layers grown during diffusion and the effect of interactions between impurities of different species.

Where diffusion is governed by the diffusion equation

$$\frac{\partial n}{\partial \theta} = \text{div grad } n \tag{1}$$

where $\theta = \int_0^t D(t)dt$ and $D = D(t)$only.

Any purely time dependent variation of the diffusion constant, due to changes in temperature, can be replaced by a change of scale in the time variable. In fact it is convenient to replace $\int D(t)dt$ by $D_1 t_1$ where D_1 is the diffusion coefficient at a standard temperature T_1 (e.g. the maximum) and t_1 is the effective extra time of diffusion at the standard temperature. In addition, the temperature variation may cause a variation in the boundary conditions, e.g. in the value of the surface concentrations. In many cases such changes will be relatively remote from the region of most interest, i.e. the p–n junction, and can be ignored. We must now evaluate this integral for cases of practical interest.

The diffusion constant $D(T)$ is given by $D_0 \exp(-E/kT)$, where E is the activation energy for diffusion. Suppose the sample is cooled at a rate a°K/sec from a temperature T°K so that

$$T = T_1 - at \tag{2}$$

The integral is then given by

$$\int_{t=0}^{t=T_1/a} D_0 \exp\{-E/k(T_1-at)\}dt \tag{3}$$

The use of zero as a bound for temperature will not cause error if diffusion has virtually ceased at the actual lower bound.

With the substitution $x = E/k(T_1-at)$ the integral becomes

$$\frac{D_0 E}{ka} \int_{x_1}^{\infty} x^{-2} e^{-x}\, dx \tag{4}$$

where $x_1 = E_1/kT_1$. This integral is related to the exponential integral.[3] For most cases of practical interest, $x_1 \ll 1$ and the integral can be repeatedly integrated by parts to give a semi-convergent series in x^{-1}, thus

$$\int_{x_1}^{\infty} x^{-2}e^{-x}dx = \frac{e^{-x_1}}{x_1^2}\left(1-\frac{2!}{x_1}+\frac{3!}{x_1^2}-\ldots\ldots\right) \tag{5}$$

Other cases may be evaluated by reference to tabulated values of the error integral[4] as

$$\int_{x_1}^{\infty} x^{-2}\, e^{-x}dx = e^{-x_1}/x_1 + Ei(-x_1) \tag{6}$$

Substituting for the integral in (4) with (5) we obtain

$$\int D dt = D_1\, t_1 \tag{7}$$

where $D_1 = D_0 \exp(-x_1)$ the diffusion coefficient at T_1 and

$$t_1 = \frac{T_1}{a}\,\frac{kT_1}{E}\left(1-\frac{2kT_1}{E}+6\left(\frac{kT_1}{E}\right)^2-\ldots\ldots\right) \tag{8}$$

An important practical case is that of substitutional impurities in silicon. For most elements

Table 1*

T_1°K	1200	1225	1250	1275	1300	1325	1350	1375	1400
ΔT	34·4	35·8	37·2	38·7	40·2	41·8	43·3	44·9	46·5
T_1°K	1425	1450	1475	1500	1525	1550	1575	1600	1625
ΔT	48·2	49·9	51·6	53·3	55·0	56·8	58·6	60·6	62·4
T_1°K	1650	1675	1700						
ΔT	64·2	66·1	68·1						

* The effective extra time t_1 for which diffusion occurs is that taken to cool through ΔT degrees when silicon cools at a constant rate from temperature T_1°K.

Vol. 2
1961

of Group III and Group V, (P, As, B, Al, Ga) we may take $E = 3{\cdot}5$ eV.[5, 6]

The range of temperature of interest for T_1 is 900–1400°C. The results of this calculation may be expressed as

$$t_1 = \Delta T/a \qquad (9)$$

where $1/a$ is the time required for the sample to be cooled through 1°, and ΔT is given in Table 1.

It is of interest to note how the higher terms of the expansion in equation (8) are most important for interstitial impurities such as copper whose activation energies are small. This may be of importance in determining the diffusion of impurities from metal contact areas.

NOTATION

a Sample cooling rate
D Diffusion coefficient of impurity
D_0 Coefficient related to D by $D = D_0 \exp(-E/kT)$
D_1 $D(T_1)$
E Activation energy of impurity in diffusion
k Boltzmann's constant
n Concentration of diffusant
t Time
t_1 Effective time of diffusion at standard (maximum) temperature
T Temperature of sample in °K
T_1 Maximum temperature
ΔT Temperature interval such that the effective extra diffusion time is given by the time for the sample to cool through ΔT°K
x E/kT
x_1 E_1/kT

Acknowledgements—The author wishes to thank the Plessey Company Ltd. for permission to publish this work, and the Ministry of Aviation for support.

B. L. H. Wilson

The Plessey Company Ltd.
Caswell, Towcester
Northants

References

1. R. A. Logan and M. Schwartz, *J. Appl. Phys.* **26,** 1287 (1955).
2. S. J. Silverman and J. B. Singleton, *J. Electrochem. Soc.* **105,** 591 (1958).
3. E. Jahnke and F. Emde, *Tables of Functions* p. 1. Dover, New York (1945)
4. *Handbook of Chemistry and Physics* p. 282. Chemical Rubber Publishing Co., Cleveland (1945).
5. W. C. Dunlap, H. V. Bohm and H. P. Mahon, *Phys. Rev.* **96,** 822 (1954).
6. C. S. Fuller and J. A. Ditzenberger, *J. Appl. Phys.* **27,** 544 (1956).

Solid-State Electronics Pergamon Press 1961. Vol. 2, p. 194. Printed in Great Britain

BOOK REVIEW

S. MROZOWSKI (Editor): **Proceedings of the Fourth Conference on Carbon.** Pergamon Press, Oxford, 1960. 778 pp., £7 10s.

THIS book contains all but five of the papers which were submitted for presentation at the Fourth Conference on Carbon held at the University of Buffalo on 15–19 June, 1959. The subject matter of the papers concerns the chemical and physical properties of graphite, electrode carbons, carbon black, activated carbons, etc.

The volume is divided into five parts having the following titles: I. Surface Properties, Adsorption and Reactivity; II. Electronic Properties; III. Carbonization, Graphitization and Structure; IV. Mechanical and Thermal Properties; V. Carbon Technology: Friction and Wear. No account is given of the discussions which followed the lectures. Author and Subject Indexes are provided.

R. MANSFIELD

Vol. 2
1961

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 195–201. Printed in Great Britain

DIRECT RECOMBINATION IN GaAs AND SOME CONSEQUENCES IN TRANSISTOR DESIGN

S. MAYBURG

General Telephone and Electronics Laboratories, Inc., Bayside 60, N.Y.

(*Received* 30 *November* 1960; *in revised form* 30 *January* 1961)

Abstract—Experiments are presented indicating that the recombination rate R in GaAs is given by the expression $R = Bpn$. The theory of Hall for direct transition is used to calculate the value of B as $1{\cdot}4\times10^{-10}$ cm^3/sec. Practical GaAs transistors are shown to be feasible if base widths of 1 μ are used.

Résumé—On présente des expériences qui indiquent que le taux de recombinaison R dans l'AsGa est donné par l'expression $R = Bpn$. La théorie de Hall pour la transition directe est utilisée pour calculer la valeur de B qui est de $1{,}4\times10^{-10}$ cm^3/sec. On démontre la possibilité de réaliser pratiquement des transistors d'AsGa, si des largeurs de base de 1 μ sont employées.

Zusammenfassung—Durch Versuche wurde nachgewiesen, dass die Rekombinationsgeschwindigkeit R in GaAs durch die Formel $R = Bpn$ darstellbar ist. Zur Berechnung des Wertes von B als $1{,}4\times10^{-10}$ cm^3/sec wurde die von Hall entwickelte Theorie der direkten Übergänge benutzt. GaAs-Transistoren sind brauchbar, wenn die Basisdicke 1 μ beträgt.

1. INTRODUCTION

RECENTLY HALL[1] summarized the theories of the recombination process in semiconductors. In particular, he discusses the radiative recombination process where the rate of recombination R is proportional to the hole density p and the electron density n with the proportionality constant B; that is

$$R = Bpn \tag{1}$$

HALL gives an expression for B in the case of direct transitions when the conduction-band minimum and the valence-band maximum occur at the same point in momentum space, a situation which, according to CALLAWAY, exists in GaAs.[2] HALL shows that

$$B_{\text{direct}} = 0{\cdot}58\times10^{-12}\mu\left(\frac{m}{m_e+m_h}\right)^{3/2}\times$$

$$\times\left(1+\frac{m}{m_e}+\frac{m}{m_h}\right)\left(\frac{300}{T}\right)^{3/2}W_g^2\ \text{cm}^3/\text{sec}$$

where W_g is expressed in electron volts and is the energy threshold for direct transition, i.e. the energy gap, m_e and m_h are the effective masses of electrons and holes, and μ is the square root of the dielectric constant. Using for the effective mass of electrons[3] $0{\cdot}08\,m$, and for the effective mass of holes[4] $0{\cdot}5\,m$, we obtain for GaAs

$$B_{\text{direct}} = 1{\cdot}4\times10^{-10}\ \text{cm}^3/\text{sec}$$

In n-type material the lifetime τ is defined in terms of the rate of recombination

$$R = p/\tau$$

and from equation (1)

$$\tau = 1/Bn$$

For GaAs doped to 10^{17}/cm^3 we would expect a lifetime of roughly $0{\cdot}1$ μsec.

Unlike GaAs, Ge and Si are indirect semiconductors. HALL[1] shows that radiative recombination by indirect transitions would yield a lifetime in Ge doped to 10^{17}/cm^3 of 150 μsec and in Si doped to 10^{17}/cm^3 of 2500 μsec. The fact that the actual lifetimes observed in Ge and Si

are much shorter than those calculated from the radiative recombination theory supports the extensive experimental data realized over the last twelve years that recombination centers determine the lifetime in Ge and Si.

In extrinsic Si or Ge, the recombination rate R is roughly proportional to the minority-carrier density and the density of recombination centers[5] N_R.
For extrinsic n-type material

$$R \sim pN_R \tag{2}$$

and the lifetime

$$\tau \sim 1/N_R$$

is roughly independent of majority-carrier density but is inversely proportional to the density of recombination centers. At high injection levels the lifetime tends to increase because a fraction of the recombination centers will not be filled at all times.

As we shall see below, in GaAs the recombination process is apparently determined by rate equation (1) and, hence, the lifetime is determined by the density of majority carriers produced by the doping or by injection. In this case lifetime will decrease with injection level. The purpose of this paper is to describe experiments which support equation (1) as describing the dominant recombination mechanism in GaAs and to discuss the device consequences of equation (1).

2. EXPERIMENTAL

If equation (1) accurately describes the recombination mechanism in GaAs, then the lifetime for holes in n-type material is

$$\tau = 1/Bn$$

where n represents all electrons present. Clearly, n is a function of injection level and therefore the lifetime is not a constant. Unfortunately, all techniques for measuring lifetime that have been developed for Ge and Si are not easily applied when the lifetime decreases with injection level. Therefore, a straightforward application of existing lifetime techniques was not used.

We have approached the problem of verifying equation (1) indirectly. Suppose we generate carriers by shining light on GaAs. Let L_1 and L_2 be the rate of generation of holes and electrons from light levels (1) and (2) respectively. Since in steady state the rate of generation of hole–electron pairs equals their rate of recombination, we have, assuming equation (1) applies

$$\left.\begin{aligned} L_1 &= R_1 = Bp_1n_1 \\ L_2 &= R_2 = Bp_2n_2 \\ \Delta L &= L_2 - L_1 = B(p_2n_2 - p_1n_1) \end{aligned}\right\} \tag{3}$$

where (p_2, n_2) and (p_1, n_1) are the hole and electron densities at levels (2) and (1), respectively. Since n_i^2 for GaAs is very small, in extrinsic n-type GaAs $p_0 = (n_i^2/n_0)$ is essentially zero, and we may write

$$n_1 = n_0 + \eta p_1$$

$$n_2 = n_0 + \eta p_2$$

where η equals 1 if there is no minority-carrier trapping, and is greater than 1 if minority trapping exists. We assume here that η is the same for all levels of illumination.

Equation (3) may then be written

$$\left.\begin{aligned} \Delta L &= B[p_2(n_0 + \eta p_2) - p_1(n_0 + \eta p_1)] \\ \Delta L &= B(p_2 - p_1)[n_0 + \eta(p_2 + p_1)] \end{aligned}\right\} \tag{4}$$

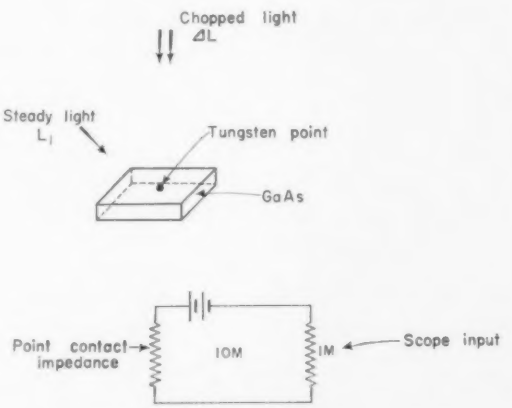


FIG. 1. Experimental arrangement. The point contact is biased to collect holes. The impedance of the point contact is large compared to the scope input impedance. Hence, any small modulation of the point contact impedance by the light will not affect the measurements. The scope measures a current which is proportional to the minority-carrier density in the surface of the slab.

or for high injection level ($n_0 \sim 10^{15}/\text{cm}^3$ in our experiments)

$$\Delta L = B\eta(p_2-p_1)(p_1+p_2) \qquad (4a)$$

Equation (4a) can be verified experimentally in GaAs in the following way. A tungsten point placed in contact with $\frac{1}{2}$-Ω-cm n-type GaAs is biased to collect holes. The current so acquired is displayed on an oscilloscope and *is assumed to be* proportional to the hole density.[6] White light of a fixed intensity and chopped at a fixed frequency of 400 c/s is shone on the GaAs in the region of the point. The chopping serves to provide a difference in light level of a fixed amount and hence a constant ΔL (see Fig. 1). When the steady light L_1 applied to the GaAs is increased, the values of p_1 and p_2 increase, but the difference p_2-p_1 decreases. In Fig. 2 we show a typical scope pattern for several levels of steady light L_1 which show the decrease in p_2-p_1 as p_1 and p_2 increase. In Table 1, data are tabulated for p_2, p_1 and $(p_2+p_1)(p_2-p_1)$ which quantitatively support equation (4a) in a limited range. At high levels of L_1 quantitative conclusions cannot be reached because of measurement errors in $|p_2-p_1|$. We have done the same experiment using a back-biased alloyed n-on-p diode instead of a point contact and have observed qualitatively the same results, although the sensitivity was so low that meaningful quantitative data could not be obtained.

In order for equation (4a) to hold, $\eta(p_2+p_1)$ must be greater than n_0. In the lowest trace in Fig. 2, p_1 is the carrier density generated by the small amount of stray light which gets by the light chopper when it is closed. Therefore, p_1 is small and could be less than $n_0 \sim 10^{15}/\text{cm}^3$. However, only ηp_2 need be greater than n_0 for equation (4a) to hold, and this appears to be the case.

In going from the lowest trace to the top trace in Fig. 2, p_2 increases by a factor of 3·5. This means that L_2 must increase by a factor of 12. The generation rate L is a product of quantum

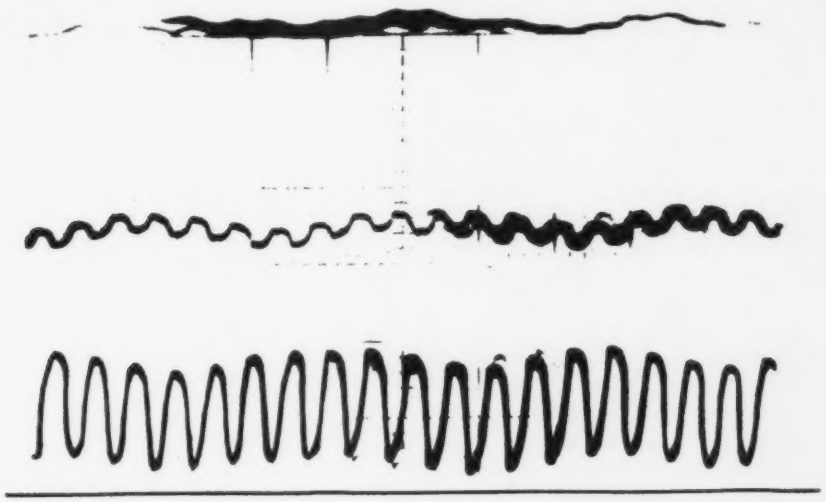


FIG. 2. The oscilloscope display of p_2 and p_1 for constant ΔL and three levels of steady illumination L_1. The base line is the trace for ΔL and L_1 equal zero. Clearly, p_2-p_1 decreases as p_2 and p_1 increase.

Table 1

p_1	p_2	$(p_2-p_1)(p_2+p_1)$
1·1	2·4	4·6
2·7	3·7	6·4
5·2	5·8	6·6

efficiency, photon flux and the absorption coefficient. White-light sources which have a high average absorption coefficient ($\mu \gg 10^3$/cm) have given larger L's than the other sources we have tried. The disadvantage of white-light sources is that it is not possible to calculate L accurately from the source characteristics. Monochromatic sources which can be more easily calibrated generally suffer from being too low in intensity to be of use. The minimum source intensity required can be estimated. Assuming $p_1 \sim n_0 = 10^{15}/\text{cm}^3$, then

$$L_1 = Bp_1(n_0+p_1) \sim 2Bn_0^2$$
$$= 3\times 10^{20}/\text{cm}^2\,\text{sec}$$

Assuming an average absorption coefficient of 10^4/cm, the photon flux must be greater than $3\times 10^{16}/\text{cm}^2$ sec to produce noticeable effects in these experiments.

The only assumption made here about the light source is that ΔL is constant during the experiment even though L_1 is increased. In order to measure B, ΔL would have to be calibrated as well as the absolute value of p_2 and p_1.

The above experiments are based on the assumption that the measured currents are proportional to the hole density. In the experiments so far described, the collector was back-biased to 22·5 V. We now wish to describe another experiment in which equation (1) is verified making use of measurements of the short-circuit current I_{sc} of the collector. We here assume, as discussed by HENISCH[6], that the short-circuit current is proportional to the hole density. At high injection level, equation (3) may be written in the steady state as

$$L = \eta Bp^2$$

We may assume that not only is $p \gg n_0$, but that $\eta \sim 1$, because the density of traps is probably not much higher than the density of donor centers, and so all traps should be filled when $p \gg n_0$. We then have

$$p \sim \sqrt{L}$$

Now, L can be varied systematically by using a point source of white light and by varying the distance r between this source and the collector–semiconductor contact. Then

$$L \sim 1/r^2$$

and if

$$p \sim I_{sc}$$

we have

$$I_{sc} \sim 1/r \qquad (5)$$

In Fig. 3 we plot I_{sc} measured using a chopped-light and a high-gain a.c. amplifier as a function of r. We see very good agreement of the experimental data with equation (5) for GaAs.

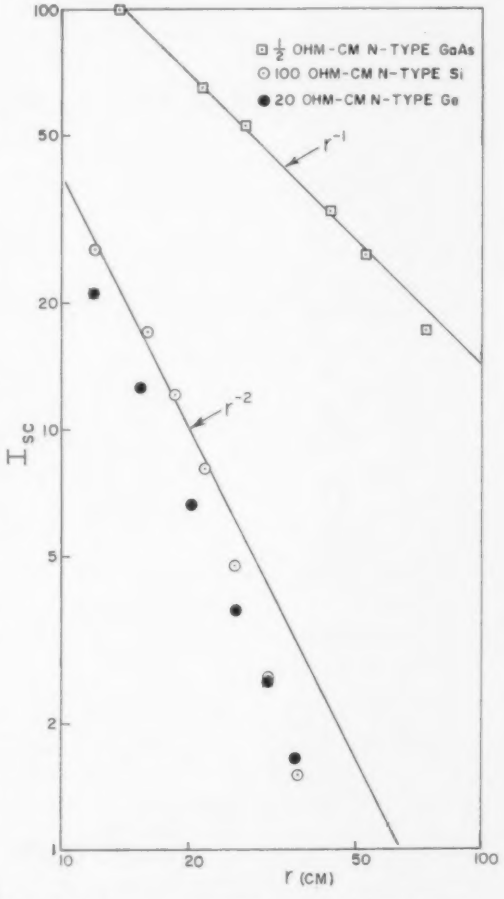


FIG. 3. Short-circuit current, I_{sc}, in arbitrary units as a function of the distance r to a constant light. The data for ½ Ω-cm n-type GaAs are shown by the squares, for 100 Ω-cm n-type Si by the open circles, and for 20 Ω-cm n-type Ge by the filled circles. For GaAs I_{sc} obeys a r^{-1} dependence, while for Ge and Si I_{sc} falls faster than r^{-2}.

Vol. 2
1961

If the rate of recombination were determined by transitions through recombination centers, then

$$\left.\begin{aligned} L &\sim p/\tau \\ p &\sim L\tau \sim \frac{1}{r^2}\tau \end{aligned}\right\} \quad (6)$$

As the rate of injection of carriers increases, the rate of recombination also increases and, at any instance, the fraction of recombination centers filled with majority carriers is reduced. Hence, τ decreases[5] as L decreases, and p will fall off faster than $1/r^2$. In Fig. 3 we plot the short-circuit current I_{sc} measured on 20 Ω-cm n-type Ge and I_{sc} measured on 100 Ω-cm n-type Si as a function of r. In agreement with equation (6), I_{sc} in both cases falls faster than $1/r^2$.

Admittedly, we have not directly verified the assumptions in our experiments that the short-circuit current or the back-biased collector current is proportional to the injected carrier density. However, the theories of point-contact behavior reviewed by HENISCH[6] indicate that these are reasonable assumptions. Furthermore, our results on GaAs, by slightly different techniques, have led to the same conclusions; in addition, our results on Ge and Si are in agreement with the way these materials are known to behave. The only way the point-contact results could be ignored would be if the point-contact current depends on the square root of the carrier density in GaAs, but depends directly on the carrier density in Ge and Si. We can see no theoretical reason why this should be so.

The analysis leading to equation (3) does not include minority-carrier diffusion effects due to non-uniform spacial generation of carriers. We have solved the diffusion equation on a computer using 10^3/cm as the value of the absorption coefficient of radiation in the GaAs. The results are qualitatively the same, and at high injection level are quantitatively in agreement with the simple analysis presented above. This can be seen intuitively: The loss of carriers by diffusion away from the surface depends on the surface concentration of carriers, but at high injection levels the loss by recombination depends on the square of the concentration. Therefore, diffusion losses become relatively less important at high injection levels.

3. CONSEQUENCES OF DIRECT RECOMBINATION IN GaAs

A. *Material purification*

As was indicated above, the lifetime in GaAs is determined by the majority-carrier density which, in turn, is determined uniquely by the doping and the injection level. Therefore, lifetime improvement by material purification, which works with germanium and silicon, may not be expected to work with GaAs. Material purification will improve mobilities, reduce traps and make crystal growing easier, but will not affect lifetime directly. The lifetime will be determined by the doping or resistivity requirement of the device used and by the injected carriers in the device during its operation. While lifetime in high-resistivity GaAs can be high, carrier injection during the lifetime measurement itself will make the apparent lifetime short.

B. *Transistor design*

Consider a diffused base n–p–n transistor with acceptor densities of 10^{19}/cm^3 in the base. From the value of B for GaAs of $1{\cdot}4\times10^{-10}$ cm^3/sec, the lifetime will be of the order of 10^{-3} μsec. Assuming a diffusion constant* for electrons D_n of 40 cm^2/sec, we get a diffusion length $\sqrt{(D_n\tau)}$ of $1{\cdot}7\,\mu$. In order to get an appreciable current gain β, a transistor must have a base width w smaller than the diffusion length.

We can estimate the emitter efficiency of such a device in the following way: The injection level n_0, at the emitter side of the base of an n–p–n transistor, which would be required to produce an emitter- to collector-current density I_{ec}, is determined from the relation

$$\left.\begin{aligned} I_{ec} &= D_n\frac{n_0}{w} \\ n_0 &= \frac{w}{D_n}I_{ec} \end{aligned}\right\} \quad (7)$$

If the doping on the base side of the emitter base junction is P and on the emitter side is N, then we can obtain the hole density p_0 on the emitter

*For the values of diffusion constants we have used measurements of Hall mobility at the corresponding carrier density under discussion. There are no drift-mobility measurements on GaAs available.

side of the junction from the relation

$$p_0 N = n_0 P$$

and, with equation (7),

$$p_0 = \frac{P}{N} n_0 = \frac{P}{N} \frac{w}{D_n} I_{ec} \tag{8}$$

The hole current in the emitter is[7]

$$I_p = p_0 \sqrt{\left(\frac{D_p}{\tau}\right)} \tag{9}$$

with D_p the hole diffusion constant in the emitter and where τ for direct recombination is

$$\tau = 1/BN \tag{10}$$

The maximum current gain β (neglecting recombination in the base layer) will be

$$\beta = I_{ec}/I_p$$

and combining equations (8)–(10) we have

$$\beta = \frac{D_n}{Pw} \sqrt{\left(\frac{N}{D_p B}\right)} \tag{11}$$

The following values apply:

$$N = 3 \times 10^{19}/\text{cm}^3$$

$$P = 10^{19}/\text{cm}^3$$

$$B = 1{\cdot}4 \times 10^{-10}\ \text{cm}^3/\text{sec}$$

$$D_n \sim 40\ \text{cm}^2/\text{sec}$$

$$D_p \sim 3\ \text{cm}^2/\text{sec}$$

$$w \sim 10^{-4}\ \text{cm}$$

Hence, $\beta = 12$. The above estimates indicate that to keep emitter efficiency high, as well as minority-carrier losses in the base low, base widths less than 1 μ are desired. Equation (11) also suggests that with lower base doping, higher emitter efficiencies would be obtained and a larger base width might be tolerated. However, to prevent "punch through", i.e. space-charge extension from the collector to the emitter, P should be no lower than $10^{18}/\text{cm}^3$. This suggests that one could lower P and raise w in such a way that the emitter efficiency would improve. In other words, if we decreased P by a factor of 10 and raised w by a factor of 3·2, β from equation (11) would increase by a factor of 3·2.

However, simultaneous with an increase in base width will come an increase in recombination current. Let us calculate the recombination current in the base of this transistor when direct recombination is dominant. From equation (1)

$$R = Bpn = B(P+n)n \tag{12}$$

where we have neglected trapping and have assumed that the added holes equal the added electrons to preserve space-charge neutrality.

The recombination current per unit area of junction I_R is given by

$$I_R = \int_{\text{emitter}}^{\text{collector}} R\,dx = \int_{n_0}^{0} R \frac{dx}{dn}\,dn$$

If we assume $dx/dn \sim w/n_0$ and that the minority-carrier density at the collector is zero we get

$$I_R = Bw\left(\frac{n_0^2}{3} + \frac{n_0 P}{2}\right) \tag{13}$$

The limitation on β from the recombination current only is obtained by combining equations (13) and (7)

$$\frac{1}{\beta} = \frac{I_R}{I_{ec}} = \frac{Bw^2}{D_n}\left(\frac{wI_{ec}}{3D_n} + \frac{P}{2}\right) \tag{14}$$

In a diffused-base transistor the high doping goes only partly through the base. If we assumed a uniform doping half way through the base, and low injection levels, then

$$\frac{I_R}{I_{ec}} = \frac{3Bw^2P}{8D_n} \tag{15}$$

The total β will be determined from both the emitter loss and the volume recombination. By combining equations (11) and (15) we have

$$\frac{1}{\beta} = \frac{I_p + I_R}{I_{ec}} = \frac{3Bw^2P}{8D_n} + \frac{Pw}{D_n}\sqrt{\left(\frac{D_p B}{N}\right)}$$

Vol. 2
1961

For $w = 3\,\mu$ and $P = 10^{18}/\text{cm}^3$ we get a β of 7, but if we reduce w to $1\,\mu$ we would have* a β of 47. Clearly, base widths close to $1\,\mu$ are desirable for a GaAs transistor.†

Acknowledgements--The author is indebted to J. BLACK and K. ARNOLD for GaAs material, and to A. CALHOUN and M. RITTERMAN for programming the diffusion equation on the computer.

*Values of β obtained obviously depend on the values of various constants used in the calculations. At high doping levels the effective masses of holes and electrons should be smaller and hence the value of B may be larger than the value used in the text. A larger B would decrease the values of β.

†In a recent paper[8] JONES and WURST of the Texas Instrument Co. described an n–p–n diffused-base GaAs transistor with β about 50. The base width was said to be very small.

REFERENCES

1. R. N. HALL, *Proc. Inst. Elect. Eng.* **106B**, Suppl. 17, 983 (1960).
2. J. CALLAWAY, *J. Electron.* **2**, 330 (1957).
3. W. G. SPITZER and J. M. WHELAN, *Phys. Rev.* **114**, 59 (1959).
4. J. T. EDMOND, R. F. BROOM and F. A. CUNNELL, *Report of the Meeting on Semiconductors, Rugby*, p. 109. Physical Society, London (1957).
5. A. F. GIBSON, *Progress in Semiconductors III* p. 139. Wiley, New York (1958).
6. H. K. HENISCH, *Rectifying Semiconductor Contacts* p. 249. Clarendon Press, Oxford (1957).
7. W. SHOCKLEY, *Electrons and Holes in Semiconductors* p. 90. Van Nostrand, New York (1950).
8. M. E. JONES and E. C. WURST, JR., GaAs *Transistors*. Fall Meeting of the AIEE, October, 1960.

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 202–208. Printed in Great Britain

A LOW-RESISTANCE OHMIC CONTACT FOR SILICON SEMICONDUCTOR DEVICES

S. L. MATLOW* and E. L. RALPH

Hoffman Electronics Corp., El Monte, California

(*Received* 2 *December* 1960)

Abstract—A low-resistance ohmic contact to lapped silicon can be obtained by a displacement plating of gold followed by an electrodeposition of copper. The contact resistance is found to be less than that obtained using unfired auto-reductive electroless nickel. The bonding strength of the contacts is found to be quite good as determined by a torque test. The mechanism of the displacement plating is discussed from the viewpoint of the factors found to influence the plating rate and adherence. A figure of merit is defined which indicates the conditions necessary to obtain good adherence. The contact resistance is found to be a linear function of the displacement plating time. A hypothesis is offered to explain these observations.

Résumé—Un contact ohmique de basse résistance jointe au silicium peut-être obtenu par un placage à déplacement d'or suivi par une superposition à électrode de cuivre. On découvre que la résistance de contact est moins que celle obtenue en utilisant un procédé autoréductif sans électrode au nickel. On démontre, avec l'aide d'une expérience à couple, que la force du lien des contacts est satisfaisante. Le mécanisme du placage à déplacement est ensuite discuté du point de vue tendant à influencer le taux de placage et l'adhérence. Un facteur de mérite indiquant les conditions nécessaires pour obtenir une bonne adhérence est défini. On démontre que la résistance de contact est une fonction linéaire de la durée de placage à déplacement. Une hypothèse est offerte pour expliquer ces observations.

Zusammenfassung—Durch elektrolytische Abscheidung von Gold und darauf folgende Elektroden-Abscheidung von Kupfer entsteht auf geläpptem Silizium ein ohmscher Kontakt mit geringem Widerstand. Der Übergangswiderstand ist kleiner als der beim Plattieren mit ungeglühtem autoreduktivem Nickel entstehende Widerstand. Die Haftfähigkeit des Kontakts wurde mittels eines Torsionsversuches geprüft und ergab sich als zufriedenstellend. Es folgt eine Diskussion der Faktoren, die die Abscheidungsgeschwindigkeit und Haftfähigkeit bei diesem Prozess beeinflussen. Die zur Erzielung guter Haftung erforderlichen Bedingungen werden durch Definition einer Leistungsziffer angegeben. Der Übergangswiderstand ist eine lineare Funktion der Abscheidungszeit. Zur Erklärung der Beobachtungen wird eine Hypothese entwickelt.

I. INTRODUCTION

DISPLACEMENT plating has been used by numerous authors to delineate *p–n* junctions in silicon. WHORISKEY used an HF solution of $Cu(NO_3)_2$.[1] SILVERMAN and BENN used a KOH solution of $KAu(CN)_2$.[2] TURNER used HF solutions of SbF_3, $CuSO_4$, $AgNO_3$, H_2PtCl_6 and $AuCl_3$.[3] He found that the gold adhered poorly. Displacement plating has also been used to make contacts to silicon. WALTZ used an HF solution of $CuSO_4$.[4] WANG used an HF and H_2SO_4 solution of $CuSO_4$.[5] DAVIES also used an HF solution of $CuSO_4$.[6] In this study an HF solution of $KAu(CN)_2$ was used.

Although the factors which determine whether displacement plating will occur are well known and have been discussed in some detail by TURNER[3], it will be necessary to discuss some aspects of displacement plating here. In order that any displacement occur on a silicon surface, it is necessary that the oxide layer be removed. HF will, of course, remove the SiO_2 by the reaction

$$SiO_2 + 6HF \rightarrow H_2SiF_6 + 3H_2O \qquad (1)$$

* Present address: Intellux, Inc., 30 S. Salsipuedes, Santa Barbara, California.

The kinetics of this reaction have been studied by BLUMBERG for the case in which the SiO_2 is in the form of vitreous silica.[7] He has found that for the temperature range of 0–35°C and for HF concentrations up to 2·5 molal, the rate is proportional to the surface area of the silica and the first power of the HF concentration. At concentrations of HF above 2·5 molal, polymerization of the HF occurs and the form of the rate law changes as a result. He has found the rate constant to be

$$k = 0{\cdot}120 \exp[-\Delta E/RT] \qquad (2)$$

g SiO_2/sec cm² HF molarity. The value of ΔE was found to be 9 kcal. Assuming that T is 300°K, the value of k is $3{\cdot}67 \times 10^{-8}$ g/sec cm² HF molarity. The densities of various forms of silica vary from about 2·20 to 2·66 g/cm³.[8] If we choose the value of 2·3 g/cm³ for the density of vitreous SiO_2, the rate constant becomes 1·60Å SiO_2/sec HF molarity. Although the structure of the oxide layer on silicon is not known, it does not seem too unreasonable to assume that it is similar to that of vitreous silica. In this study the HF concentration used was about 1·4 N. If BLUMBERG's data are at all applicable here, the rate with which we were removing the oxide layer was 2·24 Å/sec. A 500-Å layer would, therefore, take about 230 sec to be removed completely. For part of this work the silicon was in the HF solution about 600 sec. It is, therefore, not unreasonable to assume that, at plating times less than 4–6 min, the oxide was not completely removed from the surface. Since the oxide layer is irregular in thickness, some areas will be clear of oxide before others. This means that, in the first few minutes of plating, the plating will be in patches on the submacroscopic level.

There are two other factors which are known to affect the uniformity of the displacement plating. The first is the necessity for an escape path for oxidized silicon atoms. Unless the oxidized silicon atoms go into solution, the reaction will quickly stop. These escape paths or channels are present on all displacement plating. It is to fill in or cover over these channels that the gold displacement plating in this study was followed by electrolytic copper plating.

The other factor affecting the uniformity of the displacement plating on a microscopic scale is the competition between two reactions. In any given microscopic region the first few gold atoms will bond directly to silicon atoms. The subsequently plated gold atoms will bond to silicon atoms or to gold atoms. If the rate with which the gold atoms bond to silicon atoms is much greater than the rate with which gold atoms bond to gold atoms, the maximum number of gold–silicon bonds will be obtained. This would result in the bond strength of the plating being maximized. On the other hand, if the rate with which the gold atoms bond to silicon atoms is much less than the rate with which gold atoms bond to gold atoms, the minimum number of gold–silicon bonds will be obtained. This would result in the bond strength of the plating being minimized. In the latter case a large number of silicon atoms will not be bonded to gold atoms. This situation can be more easily visualized by considering a two-dimensional model. Fig. 1(a) depicts the case in which the rate with which gold

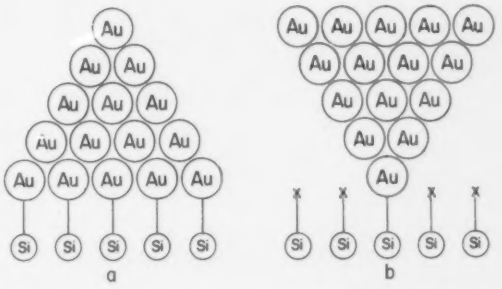


FIG. 1. Two-dimensional diagram indicating the extremes of bonding forms for the plating of gold on silicon: (a) strong bonding extreme; (b) weak bonding extreme. The x's represent unused silicon bonds. The vertical lines are covalent bonds.

bonds to silicon is much greater than the rate with which gold bonds to gold. The micro-surface of five silicon atoms is completely covered by the fifteen gold atoms, but there are five gold–silicon bonds. Fig. 1(b) depicts the other extreme case. Again the micro-surface is completely covered by the fifteen gold atoms. In this case, however, there is only one gold–silicon bond and four "dangling" silicon bonds. The "dangling" bonds will never be plated since the pyramid of this micro-surface will bridge with the pyramids of adjacent micro-surfaces and the silicon will no longer "see" the plating solution.

As TURNER[3] has pointed out, displacement

plating is actually a microscopic electrolytic cell reaction. This system can be considered as the composite of six "reactions": (A) the reaction between the gold and the silicon or the gold and gold, (B) the flow of electrons or holes or both through the silicon, (C) the reaction between silicon and the fluoride ions, (D) the diffusion of fluoride ions to the silicon surface, (E) the diffusion of ions in the bulk of the solution to retain electrical neutrality in the solution, and (F) the diffusion of gold ions to the silicon surface. The displacement plating will occur as fast as the slowest of these reactions. When the concentrations of the reactants are very high as in the case of HF–HNO_3 etching [in this case (A) and (F) refer to HNO_3 rather than gold ions], (D), (E) and (F) could be expected to be the limiting reactions. ROBBINS and SCHWARTZ have found this to be the case.[9] They found that the over-all etching rate was diffusion limited. Unless the resistivity of the silicon is high, it is not expected that (B) would be the rate limiting reaction.*

The pertinent electrodes for the system studies are:[10]

$$Si + 6F^- \rightleftharpoons SiF_6^= + 4e^-, \quad E^0 = -1{\cdot}2\ V \tag{3}$$

$$Au + 2CN^- \rightleftharpoons Au(CN)_2^- + e^-, \quad E^0 = -0{\cdot}6\ V \tag{4}$$

$$Au \rightleftharpoons Au^+ + e^-, \quad E^0 = \text{ca.} + 1{\cdot}68\ V \tag{5}$$

The combination of (4) and (5) yield an association constant for the gold cyanide complex of $3{\cdot}98 \times 10^{38}$. The equilibrium constant for the reaction

$$HCN \rightleftharpoons H^+ + CN^- \tag{6}$$

is 4×10^{-10}.[11] Thus, the equilibrium constant for the reaction

$$Au^+ + 2HCN \rightleftharpoons Au(CN)_2^- + 2H^+ \tag{7}$$

is

$$6{\cdot}37 \times 10^{19} = \frac{[H^+]^2[Au(CN)_2]}{[Au^+][HCN]^2} \tag{8}$$

or, solving for $[Au^+]$,

$$[Au^+] = \frac{[H^+]^2[Au(CN)_2]}{6{\cdot}37 \times 10^{19}[HCN]^2} \tag{9}$$

A solution made by dissolving 1 g of $KAu(CN)_2$ (41 per cent Au) in 100 ml of H_2O, to which 5 ml HF (48 per cent) is added, has the approximate concentrations: $KAu(CN)_2$, $1{\cdot}98 \times 10^{-2}$ M; KCN, $6{\cdot}45 \times 10^{-3}$ M; HF, 1·36 M. If the KCN is assumed to be completely converted to HCN by the HF, the value $6{\cdot}45 \times 10^{-3}$ can be used for [HCN] in equation (9). Substituting these various values into (9) we find that

$$[Au^+] = 1{\cdot}38 \times 10^{-17}\ M \tag{10}$$

Thus, it is rather unlikely, although not impossible, that Au^+ is the plating species rather than $Au(CN)_2^-$. Since (4) is apparently the appropriate electrode, one would expect hydrogen evolution. A gas was observed to be liberated during our plating studies. However, no attempt was made to analyze the gas.

II. PLATING ON *n*-TYPE SILICON

The initial experimental work was done on 0·5 Ω-cm *n*-type silicon wafers 1 cm × 2 cm × 0·05 cm. The gold plating solution used was made up by dissolving 1 g of $KAu(CN)_2$ (41 per cent Au) in 100 ml of water to which was added 5 ml (48 per cent) HF. The first wafers used were chemically polished. The gold film adhered poorly, as was demonstrated by the flaking which occurred when the plating was scratched with the tips of a pair of tweezers. The solution was then tried on lapped wafers using different grades of abrasives. The wafers lapped with very fine abrasives did not yield well-bonded plating. When no. 600 abrasive was used, the plating adhered satisfactorily.

It was found that, when the wafers were plated in the dark, the plating was thinner than that obtained by plating under room illumination conditions. To test the electrical characteristics of the plating, wafers were plated for various times under room illumination and then copper electroplated onto the gold for 5 min at 25 mA/cm². As pointed out above, this was done to fill up or cover over the channels in the gold. The copper solution was made by dissolving 21 g of $CuSO_4 \cdot 5H_2O$ in

* It has been pointed out to the authors that, even though the resistivity of the silicon is high, (B) will not be rate-limiting if the surface recombination velocity is high.

Vol. 2
1961

100 ml of water, to which was added 2·9 ml of concentrated H_2SO_4.

After plating, the 1 cm × 2 cm faces of the wafers were masked with Mylar tape and edges etched to separate the faces electrically. The I–V characteristics of the plated wafers were plotted on an x–y plotter. A given wafer was placed in series with a known resistance and the voltage drop across the wafer and the voltage drop across the resistance were measured with the x–y plotter. In every case the plots were found to be linear indicating that, within the precision of the measurements, the contacts were ohmic.

In Fig. 2 are plotted the resistance measured for various gold-plating times, before and after solder dipping in 60/40 Sn–Pb solder. As the plating time was decreased, the total resistance of the system decreased. When the gold-plating time was decreased to less than 10 min the bonding became so poor that the copper films did not adhere well. This plating time corresponded to a combined resistance of 0·2 Ω for the two contacts and the bulk silicon. The hot electroless nickel-plating technique[12] gave a resistance of 1·0 Ω. It should be pointed out that with neither the Au–Cu nor with the Ni plating were the samples "annealed". Thus, the Ni resistances are considerably higher than those obtained by SULLIVAN and EIGLER after heat treating the plated silicon.[12]

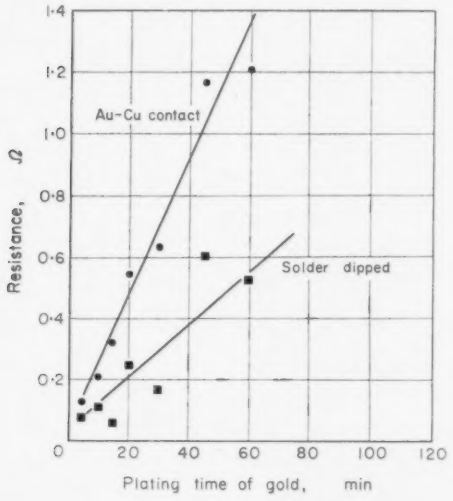


FIG. 2. Total resistance vs. plating time of gold–copper contact on 1 × 2 × 0·05 cm silicon blanks. (0·5 Ω-cm, n-type silicon, before and after solder dipping.)

According to the contact-resistance work of WURST and BORNEMAN, gold should give a rectifying contact on n-type silicon.[13] Their theoretical treatment was based upon work functions. GORDY and THOMAS have found that a correlation exists between work function and electronegativity.[14] Their relation is

$$X = 0{\cdot}44\Phi - 0{\cdot}15 \qquad (11)$$

where X is the electronegativity and Φ is the work function. Since this is the case, one can discuss barrier layers either in terms of work function or in terms of electronegativity. One of the difficulties in using work functions is the problem of deciding whose value to use. For example, MICHAELSON's compilation of work functions gives values for the work function of gold from 4·0 to 4·92 eV.[15] Unfortunately the same situation prevails with respect to electronegativities. ALLRED and ROCHOW list twelve different values of the electronegativity of silicon ranging from 1·64 to 1·99.[16] In Table 1 are listed the electronegativity values reported by several workers. It is clear that the decision as to which values to use is, at best, a difficult one. What we have chosen to do (and it may well be the

Table 1. Selected values of electronegativity

Element	Ref. 17	Ref. 18	Ref. 19	Ref. 20	Ref. 21	Average
Si	1·74	1·82	1·84	1·8	1·90	1·82
Cu	1·75	0·96	1·07	1·8*	1·34	1·38
Au	†	†	†	2·3*	2·87	2·58
Ni	1·75	†	†	1·7*	2·13	1·86

* These values were obtained by the Pauling method by HAISSINSKY[22].
† No value given for this element.

wrong choice) is to use the average of the values of Table 1.

Since the electronegativity values for silicon are based upon calculations on the atom or simple molecules, the value should be associated with intrinsic silicon. Since the work function of *n*-type silicon is about 0·5 eV less than that of intrinsic silicon, the electronegativity should be about 0·22 less [equation (11)]. This would make the electronegativity of *n*-type silicon about 1·60.

It will not be possible, here, to discuss the quantitative theory of barrier-layer rectification in the electronegativity language. The general result, however, is that if the electronegativity of the metal is greater than 1·82 it will be rectifying on *n*-type silicon. The "degree of rectification" will be greater, the greater is the difference between the two electronegativities. If the electronegativity of the metal is less than 1·82 but greater than 1·60, the contact will be ohmic but of high resistance. The resistance will drop as the metal electronegativity approaches 1·60. Thus, gold should be rectifying on *n*-type silicon. If, however, the gold layer is quite thin and is plated with copper, electrons from the copper will "spill" over into the gold-raising Fermi level and thus lowering the electronegativity. As the gold layer increases in thickness the "spillage" from the copper will have less of an effect on the gold electronegativity. Thus, the contact will have a higher resistance. This, we believe, is the explanation for the contact not being rectifying and the explanation for the curves of Fig. 2.

III. PLATING ON SOLAR CELLS

Solar cells were also plated with our gold–copper system. The solar cells were 1 cm × 2 cm × 0·05 cm. It was found that, although the *n*-side plated quite well, the *p*-side would not plate well. We then tried plating the *p*-side in the dark. It was found that the rate of plating was much faster in the dark; however, the adherence of the plating was poor. The dark-plating of the *p*-side was repeated using lower concentrations of gold and it was found that the plating rate decreased *but the adherence improved*. It was found that 10 min in the dark in a solution of 0·1 g $KAu(CN)_2$ (41 per cent Au) in 100 ml H_2O to which 5 ml of HF (48 per cent) was added, gave a satisfactory plating rate and adhered quite well.

A series of solar cells was plated by using the concentrated gold solution in the light to plate the *n*-sides and the diluted solution in the dark to plate the *p*-sides. The cells were then copper plated by the method discussed above and subsequently solder dipped. These solar cells were compared with cells plated with the hot electroless nickel system.[12] The nickel was not heat treated since that would have resulted in the shorting of the junction. The dark *I–V* characteristics of the cells were measured and the slope of the curve, at a forward current of 100 mA, was used as the measurement of comparison. This measurement was not the true device series resistance since the forward current was not great enough to be in the series-resistance-limited portion of the curve. However, this apparent series-resistance measure-

Table 2. Comparison of solar-cell contacts

Cell no.	Apparent series resistance (Ω)	
	Gold–copper contact	Electroless nickel contact
1	1·50	1·75
2	1·25	1·75
3	1·20	1·50
4	0·75	1·60
5	0·75	1·50
6	0·75	1·50
7	1·10	1·50
8	1·00	1·50
9	1·00	1·50
10	1·00	1·25
11	1·00	1·25
12	1·00	1·25
13	1·00	1·75
14	1·00	1·60
15	1·00	1·30
16	1·50	1·50
17	1·40	1·50
18	1·25	1·50
19	1·20	1·75
20	1·20	1·75
21	1·20	1·75
22	1·50	2·00
23	1·50	1·75
24	1·25	1·75
25	1·50	1·75
26	1·25	1·75
27	1·25	1·50
Mean	1·20 ± 0·16	1·60 ± 0·19

ment was very useful for comparing the merits of these contacts. The results are listed in Table 2.

The gold–copper contact had a mean series resistance of 1·20 Ω with a mean deviation of 0·16 Ω, while the nickel contact had a mean series resistance of 1·60 Ω with a mean deviation of 0·19 Ω. This corresponded to a 90 per cent confidence interval of 1·09–1·23 Ω for the gold–copper contact and 1·52–1·64 Ω for the nickel contact.

A mechanical test was devised to test the bond strength of the contacts. A sketch of the test apparatus is shown in Fig. 3. Weights were applied on a shingled pair of cells so that a torque was

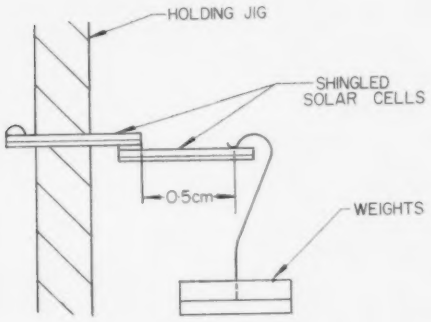


FIG. 3. Mechanical-bond-strength test apparatus.

applied to the contact, thus tending to peel off the plating. The gold–copper contacts and the nickel contacts both held 715 g without pulling loose. Larger weights were not applied because previous tests had shown that very little more weight was required to break the 0·05-cm-thick solar cells.

After a one-month shelf test, the samples were retested and none of the cells showed a series resistance increase of more than 0·25 Ω; none showed any decrease in bonding strength.

IV. THEORY

WHORISKEY's $Cu(NO_3)_2$–HF system plates faster on *n*- than on *p*-type silicon.[1] The plating is influenced by illumination. The lower the resistivity of the *n*-region, the faster is the rate of deposition. His observations are in accord with the present work. Unfortunately, he does not report what the concentrations of his substituents were. SILVERMAN and BENN, using $KAu(CN)_2$ and KOH, reported essentially the same observations.[2] Their concentrations were $3{\cdot}47\times10^{-2}$ M $KAu(CN)_2$ and 3·56 M KOH. On the other hand, ILES and COPPEN used HF solutions of AgF and $CuSO_4$.[23] They found that, regardless of whether the samples were in the light or the dark, the plating always occurred predominantly on the side of lower conductivity. On diluting their solutions they found that illumination affected the plating rate and that the *n*-side always plated faster. They do not give the concentrations of their solutions. For the AgF–HF situation, however, they do state that they used a 10:1 dilution of an aqueous solution of AgF saturated at 0°C. To 5 ml of the diluted solution was added one drop of HF. If we assume that their saturated solution was 6·76 M AgF,[24] and that one drop is 0·05 ml, their final solution was 0·614 M AgF and $2{\cdot}86\times10^{-2}$ M HF. TURNER[3] obtained results "intermediate" between two extremes. His concentrations were 2·24 M SbF_3 and 14·3 M HF; 0·801 M $CuSO_4\cdot5H_2O$ and 0·286 M HF; 0·235 M $AgNO_3$ and 0·114 M HF; $5{\cdot}79\times10^{-3}$ M $H_2PtCl_4\cdot6H_2O$ and $5{\cdot}72\times10^{-2}$ M HF; and $3{\cdot}295\times10^{-2}$ M $AuCl_3$ and $8{\cdot}58\times10^{-3}$ M HF. Our solution, as indicated above, was $1{\cdot}98\times10^{-2}$ M $KAu(CN)_2$ and 1·36 M HF.

It would appear from these data that, when the concentration of the metal ion is two orders of magnitude smaller than the HF or KOH concentration, reaction (A) of the Introduction is the rate-limiting step. The reaction is affected by light and is faster on *n*- than on *p*-type silicon. When the concentration of the metal ion is two orders of magnitude greater than the HF concentration, (C) is the rate-limiting step. The rate is not affected by illumination and is independent of whether the silicon is *n*- or *p*-type. When the two concentrations are within an order of magnitude, neither reaction is definitely rate limiting and effects of both steps are manifest. Thus, from our data and the data cited above, we conclude that reaction (A) is

$$Si^x + M^+ + n \xrightarrow{k_1} Si{-}M \qquad (12)$$

where Si^x is a surface silicon atom with a single electron in the "dangling" bond (nothing being

said about the status of the other three silicon orbitals), M^+ is Au^+ or $Au(CN)_2^-$ or something "in between", and n is a conduction-band electron. Furthermore, it was concluded that reaction (C) is

$$Si^x + F^- \xrightarrow{k_2} Si{-}F + n \qquad (13)$$

or similar reactions which add further fluorines to the silicon atoms, leading to $SiF_6^=$. The role which light plays is to produce a photocurrent, injecting electrons into the n-side of the junction and holes into the p-side. The excess holes recombine with conduction electrons, further reducing their concentration on the p-side.

The effects of the various factors on the bond strength of the plating still remain to be explained. These are, as shown in Section III, that the bond strength increased with increasing concentration of conduction-band electrons, with increasing "roughness" of the surface and with decreasing gold concentration. In light of what was said regarding Fig. 1, these effects can be explained as follows. If the gold-on-gold plating portion of reaction (A) is

$$Si{-}Au + 2Au^+ \xrightarrow{k_3} \begin{matrix} Au \\ \vdots \\ Si{-}Au \\ \vdots \\ Au \end{matrix} + 2p \qquad (14)$$

where the dotted lines represent metallic, rather than covalent, bonding, the bond strength of the plating is a direct (rather than inverse) function of a figure of merit f. The figure of merit f is defined as the ratio of the rate of reaction (12) to the rate of reaction (14) at a time when an appreciable fraction of the micro-surface is not yet covered with gold atoms. If (s) is the concentration of surface sites per cm^2 of projected surface,

$$f = \frac{k_1(s)(n)}{k_3(Si{-}Au)(M^+)} \qquad (15)$$

It is clear that the form of f explains the effects mentioned above.

REFERENCES

1. P. J. Whoriskey, *J. Appl. Phys.* **29,** 867 (1958).
2. S. J. Silverman and D. R. Benn, *J. Electrochem. Soc.* **105,** 170 (1958).
3. D. R. Turner, *J. Electrochem. Soc.* **106,** 701 (1959).
4. M. C. Waltz, *U.S. Pat.* 2814589 (1957).
5. P. Wang, Recent News Paper, *The Electrochemical Society Meeting, New York* (1958).
6. L. W. Davies, *J. Sci. Instrum.* **36,** 423 (1958).
7. A. A. Blumberg, *J. Phys. Chem.* **63,** 1129 (1959).
8. C. D. Hodgman, *Handbook of Chemistry and Physics* (40th Ed.) p. 646–7. Chemical Rubber Publishing Co., Cleveland (1958).
9. H. Robbins and B. Schwartz, *J. Electrochem. Soc.* **106,** 505 (1959).
10. W. M. Latimer and J. H. Hildebrand, *Reference Book of Inorganic Chemistry* (3rd Ed.) pp. 125 and 528. MacMillan, New York (1951).
11. W. M. Latimer and J. H. Hildebrand, *Reference Book of Inorganic Chemistry* (3rd Ed.) p. 562. MacMillan, New York (1951).
12. M. V. Sullivan and J. H. Eigler, *J. Electrochem. Soc.* **104,** 226 (1957).
13. E. C. Wurst, Jr. and E. H. Borneman, *J. Appl. Phys.* **28,** 235 (1957).
14. W. Gordy and W. J. O. Thomas, *J. Chem. Phys.* **24,** 439 (1956).
15. H. B. Michaelson, *Handbook of Chemistry and Physics* (40th Ed.) (Edited by C. D. Hodgman) p. 2554. Chemical Rubber Publishing Co., Cleveland (1958).
16. A. L. Allred and E. G. Rochow, *J. Inorg. Nucl. Chem.* **5,** 269 (1958).
17. A. L. Allred and E. G. Rochow, *J. Inorg. Nucl. Chem.* **5,** 264 (1958).
18. W. Gordy, *Phys. Rev.* **69,** 604 (1946).
19. H. O. Pritchard and H. A. Skinner, *Chem. Rev.* **55,** 745 (1955).
20. L. Pauling, *The Nature of the Chemical Bond* p. 60 (2nd Ed.). Cornell University Press, Ithaca (1948).
21. M. A. Fineman and R. Daignault, *J. Inorg. Nucl. Chem.* **10,** 205 (1959).
22. M. Haissinsky, *J. Phys. Radium* **7,** 7 (1946).
23. P. A. Iles and P. J. Coppen, *J. Appl. Phys.* **29,** 1514 (1958).
24. W. F. Linke and A. Seidell, *Solubilities of Inorganic and Metal–Organic Compounds* (4th Ed.) p. 90. Van Nostrand, Princeton (1958).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 209–215. Printed in Great Britain

THE EVALUATION OF "ESAKI INTEGRALS" AND AN APPROXIMATE EXPRESSION FOR THE TUNNEL-DIODE CHARACTERISTIC

T. P. BRODY and R. H. BOYER*

Westinghouse Research Laboratories, Pittsburgh 35, Pennsylvania

(*Received* 17 *October* 1960; *in revised form* 18 *November* 1960)

Abstract—Fermi–Dirac-type integrals, involving a product of two density-of-states functions in the integrand, have been first considered by Esaki. Two series expressions, suitable for computation, have been derived for such integrals, appropriate for room temperature and low temperatures respectively. A simple analytical current–voltage characteristic for the tunnel diode has been obtained, which has approximately the same degree of validity as Esaki's original theory.

Résumé—Les intégrales du type Fermi–Dirac, qui comprennent un produit de deux fonctions de densité de constitution, ont été considérées tout d'abord par Esaki. Deux formules en série, adaptées au calcul, ont été dérivées pour de telles intégrales à la température ambiante et aux basses températures respectivement. Une simple courbe analytique courant–tension pour la diode tunnel a été obtenue ayant environ le même degré de validité que la première théorie d'Esaki.

Zusammenfassung—Esaki hat als erster Fermi–Dirac-Integrale untersucht, die das Produkt von zwei Verteilungsfunktionen der Energiezustände enthalten. Für diese Integrale wurden zwei Reihenentwicklungen abgeleitet, die sich für Berechnungen bei Zimmertemperatur und bei niedrigen Temperaturen eignen. Es ergab sich eine einfache analytische Strom–Spannung-Kennlinie für die Tunneldiode, die etwa denselben Gültigkeitsgrad besitzt wie Esakis ursprüngliche Theorie.

INTRODUCTION

In the theory of the tunnel diode, as originally proposed by Esaki[1], integrals of the form

$$\int_{E_C}^{E_V} f(E)\rho_C(E)\rho_V(E)dE$$

occur, where $f(E)$ = Fermi–Dirac function; ρ_C, ρ_V = density-of-states functions; E_V, E_C = valence, conduction band edges. While there is a good deal of doubt regarding the appropriate form of the density-of-states functions for the tunnel-diode problem[2],† we have thought it worth while to evaluate the integral analytically for the case of "parabolic bands" considered by Esaki, i.e. for functions of the form

$$\rho_C(E) = C'(E-E_C)^{1/2}$$
$$\rho_V(E) = C''(E_V-E)^{1/2}$$
$$C', C'' = \text{constants}$$

these functions being appropriate to a non-perturbed band structure.‡ It turns out that the

* Now at the Department of Mathematics, McGill University, Montreal, Quebec, Canada.

† *Note added in proof:* It is implied by recent theoretical work [e.g. E. O. Kane, *J. Appl. Phys.* **32**, 83 (1961)] that density-of-states considerations are not of importance for the major part of the tunnel characteristic. We would prefer not to comment on this at present except to point out that the situation is in many ways obscure and subject to disagreement, e.g. Kane's "Stark ladder" density-of-states assumption was criticized by Price [*IBM J. Res. Dev.* **3**, 364 (1959)].

‡ An approximate method of evaluating Esaki integrals for other types of density-of-states functions, such as power-series and Gaussian type, has been developed by one of the authors (T.P.B.)[3]. Such functions may be more appropriate for characterizing a band structure which is heavily perturbed due to a large density of impurity states.

series obtained on evaluating the integral suggests a simple expression for the current–voltage characteristic of a tunnel diode, which reproduces the experimental characteristic over the major portion of the tunneling region, and which may be found useful in the analysis of circuits containing tunnel diodes.

Two methods of evaluating the integral will be given. The first one is appropriate to the situation at room temperature, but its convergence becomes poor at low temperatures, when the second method has to be used.

ROOM-TEMPERATURE EXPANSION

The integral to be evaluated then has the form ($\beta = 1/kT$)

$$I = \int_{E_C}^{E_V} \frac{(E_V - E)^{1/2}(E - E_C)^{1/2}}{\exp[\beta(E - E_F)] + 1} dE \qquad (1)$$

In the first method we consider instead a complex integral around the contour C in Fig. 1,

$$I' = \int_C \frac{(E - E_V)^{1/2}\,(E - E_C)^{1/2}}{\exp\,[\beta E(-E_F)] + 1} dE$$

where the denominator is taken to be positive for large positive E. We can see that $I = \frac{i}{2} I'$. The integrand of I' has poles when

$\exp\,[\beta(E - E_F)] + 1 = 0$, i.e. at $E = E_F + \dfrac{\pi n i}{\beta} \equiv E_n$ ($n = \pm 1, \pm 3, \ldots$).

Now

$$(E - E_V)^{1/2}\,(E - E_C)^{1/2} =$$

$$= E\left(1 - \frac{E_V + E_C}{E} + \frac{E_V E_C}{E^2}\right)^{1/2}$$

$$= E\left[1 - \frac{E_V + E_C}{2E} + \frac{E_V E_C}{2E^2} - \frac{1}{8}\frac{(E_V + E_C)^2}{E^2} + O\left(\frac{1}{E^3}\right)\right]$$

$$= E - \frac{E_V + E_C}{2} - \frac{(E_V - E_C)^2}{8E} + O\left(\frac{1}{E^2}\right)$$

as $E \to \infty$, so

$$I = \frac{i}{2}\int_C \frac{(E - E_V)^{1/2}(E - E_C)^{1/2} - [E - (E_V + E_C)/2 - (E_V - E_C)^2/8E]}{\exp\,[\beta(E - E_F)] + 1} dE \qquad (2)$$

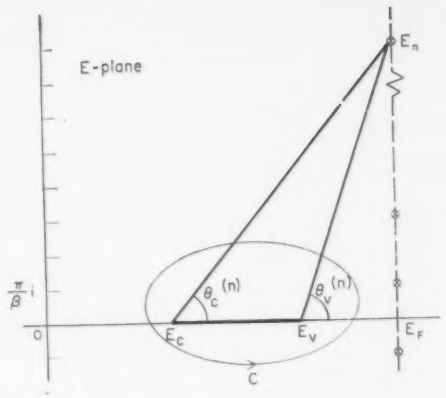


FIG. 1.

as the second term in the numerator has no poles within C and thus does not contribute to $\int_C$. The numerator is now $O(1/E^2)$ as $E \to \infty$, and the denominator is bounded away from zero on the sequence of squares with corners at

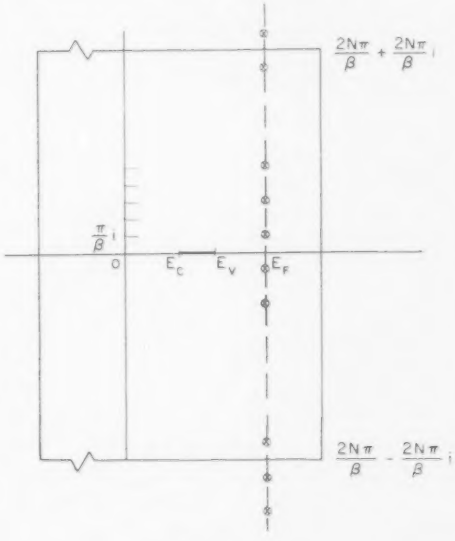


FIG. 2.

Vol. 2
1961

$$\left(\pm\frac{2N\pi}{\beta}\pm\frac{2N\pi}{\beta}i\right);\ (N=1,2,\ldots)$$

shown in Fig. 2. Thus it follows from the theory of residues that

$$I=\frac{i}{2}(-2\pi i)\sum_{n\ \text{odd}}\text{Res}\,F(E)_{E=E_n}+$$

$$+\frac{i}{2}(-2\pi i)\,\text{Res}\,F(E)_{E=0}$$

Here $\tan\theta_j^{(n)}=\dfrac{\pi n}{\beta(E_F-E_j)}$

$$(0<\theta_j^{(n)}<\pi \text{ and } \theta_j^{(n)}\to\frac{\pi}{2} \text{ as } n\to\infty)$$

Simplifying and using the identity

$$\frac{1}{1+e^{-x}}\equiv\frac{1}{2}+2x\sum_{n>0,\ \text{odd}}\frac{1}{x^2+\pi^2n^2}$$

we get

$$I=\frac{\pi}{16}(E_V-E_C)^2+\frac{2\pi}{\beta}\sum_{n>0,\ \text{odd}}\left\{\left[E_F-\frac{E_V+E_C}{2}\right]\right.$$

$$\left.-\cos\left(\frac{\theta_V^{(n)}}{2}+\frac{\theta_C^{(n)}}{2}\right)\left[(E_F-E_V)^2+\frac{\pi^2n^2}{\beta^2}\right]^{1/4}\left[(E_F-E_C)^2+\frac{\pi^2n^2}{\beta^2}\right]^{1/4}\right\} \tag{3}$$

where $F(E)$ denotes the integrand in (2).

For E near E_n

$$\exp[\beta(E-E_F)]+1=a_1^{(n)}(E-E_n)+a_2^{(n)}(E-E_n)^2+\ldots$$

To find $a_1^{(n)}$, we differentiate and set $E=E_n$:

$$\beta\exp[\beta(E_n-E_F)]=-\beta=a_1^{(n)}$$

independently of n, so

$$I=\frac{\pi}{\beta}\sum_{n\ \text{odd}}\left\{\left[E_n-\frac{E_V+E_C}{2}-\frac{(E_V-E_C)^2}{8E_n}\right]\right.$$

$$\left.-(E_n-E_V)^{1/2}(E_n-E_C)^{1/2}\right\}+\frac{\pi}{8}\frac{(E_V-E_C)^2}{\exp(-\beta E_F)+1}$$

To evaluate the sum, we take the residues in conjugate pairs. It is more convenient to introduce the angles $\theta_C^{(n)}$ and $\theta_V^{(n)}$ (see Fig. 1), and in terms of these we get

$$I=\frac{2\pi}{\beta}\sum_{n>0,\ \text{odd}}\left\{\left[E_F\frac{E_V+E_C}{2}-\frac{(E_V-E_C)^2}{8}\cdot\frac{E_F}{E_F^2+\frac{\pi^2n^2}{\beta^2}}\right]\right.$$

$$\left.-\left[(E_F-E_V)^2+\frac{\pi^2n^2}{\beta^2}\right]^{1/4}\left[(E_F-E_C)^2+\frac{\pi^2n^2}{\beta^2}\right]^{1/4}\cos\left(\frac{\theta_V^{(n)}}{2}+\frac{\theta_C^{(n)}}{2}\right)\right\}+\frac{\pi}{8}\frac{(E_V-E_C)^2}{\exp(-\beta E_F)+1}$$

The series converges with at least the rapidity of

$$\sum_{n>0,\ \text{odd}}\frac{1}{n^2}$$

however, for βE_F large, many terms will be required. In particular, when

$$E_F=\frac{E_V+E_C}{2}$$

the series vanishes (since the cosine vanishes) and

$$I=\frac{\pi}{16}(E_V-E_C)^2$$

independently of β.

If we now consider the total Esaki current

$$I_T=I_C-I_V=A\int_{E_C}^{E_V}(f_C-f_V)(E_V-E)^{1/2}(E-E_C)^{1/2}dE$$

o

where $A = C'C''Z$ and Z = tunneling probability (assumed constant over the range of integration) and write

$$\begin{aligned}\beta E &= x\\ \beta E_V &= x_V\\ \beta E_C &= x_C\\ \beta E_{F_C} &= \eta\\ \beta E_{F_V} &= \xi\end{aligned}$$

we get

$$I_T = I_C - I_V = A\,\frac{2\pi}{\beta^2}\sum_{n>0,\,\text{odd}}\left\{[\eta-\xi]+\cos_V\left(\frac{\theta_C^{(n)}}{2}+\frac{\theta_V^{(n)}}{2}\right)[(\xi-x_V)^2+\pi^2n^2]^{1/4}[(\xi-x_C)^2+\pi^2n^2]^{1/4}\right.$$
$$\left.-\cos_C\left(\frac{\theta_C^{(n)}}{2}+\frac{\theta_V^{(n)}}{2}\right)[(\eta-x_V)^2+\pi^2n^2]^{1/4}[(\eta-x_C)^2+\pi^2n^2]^{1/4}\right\} \qquad (4)$$

Here $\cos_V$ and $\cos_C$ denote the cosine factors corresponding to E_{F_V} and E_{F_C} on the complex E-plane. These terms can be transformed, using the identity

$$2\cos^2\frac{A+B}{2} = 1+\cos(A+B)$$

Thus

$$\cos\frac{\theta_C+\theta_V}{2} = \sqrt{\frac{1+\cos\theta_C\cos\theta_V-\sin\theta_C\sin\theta_V}{2}}$$

From Fig. 1 we have

$$\cos_C\theta_V^{(n)} = \frac{\eta-x_V}{[(\eta-x_V)^2+n^2\pi^2]^{1/2}}$$

$$\cos_C\theta_C^{(n)} = \frac{\eta-x_C}{[(\eta-x_C)^2+n^2\pi^2]^{1/2}}$$

and similar expressions for $\cos_V$.
Also

$$\sin_C\theta_V^{(n)} = \frac{n\pi}{[(\eta-x_V)^2+n^2\pi^2]^{1/2}}$$

$$\sin_C\theta_C^{(n)} = \frac{n\pi}{[(\eta-x_C)^2+n^2\pi^2]^{1/2}}$$

and similarly for $\sin_V$.

Then

$$\cos_C\left(\frac{\theta_C^{(n)}+\theta_V^{(n)}}{2}\right) = \frac{1}{\sqrt{2}}\left\{1+\right.$$
$$\left.+\frac{(\eta-x_V)(\eta-x_C)-n^2\pi^2}{[(\eta-x_V)^2+n^2\pi^2]^{1/2}[(\eta-x_C)^2+n^2\pi^2]^{1/2}}\right\}^{1/2}$$

and similarly for $\cos_V$.

If these values are substituted in equation (4), after simplification, we get, using

$$\eta_C = \eta-x_C \qquad \eta_V = \eta-x_V$$

$$\xi_C = \xi-x_C \qquad \xi_V = \xi-x_V,$$

$$I_T = \frac{2\pi A}{\beta^2}\sum_{n>0,\,\text{odd}}\left\{(\eta-\xi)-\frac{1}{\sqrt{2}}[(\eta_C^2\right.$$
$$+n^2\pi^2)^{1/2}(\eta_V^2+n^2\pi^2)^{1/2}+\eta_V\eta_C-n^2\pi^2]^{1/2}$$
$$+\frac{1}{\sqrt{2}}[(\xi_C^2+n^2\pi^2)^{1/2}(\xi_V^2+n^2\pi^2)^{1/2}$$
$$\left.+\xi_C\xi_V-n^2\pi^2]^{1/2}\right\} \qquad (5)$$

The ambiguity of the sign of the $[\ ]^{1/2}$ term is resolved by inspection of Fig. 1. We see that when η, ξ are $<(x_C+x_V/2)$, then $(\theta_C+\theta_V)/2>\pi/2$, so the negative root has to be taken. However, if $\eta, \xi>(x_C+x_V)/2$, $\cos(\theta_C+\theta_V)/2>0$.

For the symmetrical case $E_V-E_{F_V} = E_{F_C}-E_C$, we find that η is always $\geqslant(x_C+x_V)/2$ while $\xi\leqslant(x_C+x_V)/2$ for all forward voltages ($\eta>\xi$), and the expression then simplifies to

$$I_T = \frac{2\pi A}{\beta^2}\sum_{n>0,\,\text{odd}}\left\{(\eta-\xi)-\sqrt{2}[(\eta_C^2+n^2\pi^2)^{1/2}\right.$$
$$\left.\times(\eta_V^2+n^2\pi^2)^{1/2}+\eta_V\eta_C-n^2\pi^2]^{1/2}\right\} \qquad (5a)$$

Vol. 2
1961

It can easily be shown that the $\sqrt{2}[\ ]^{1/2}$ term reduces to

$$\left[(\eta_C+\eta_V)^2-\frac{(\eta_C^2-\eta_V^2)^2}{4n^2\pi^2}+O\left(\frac{1}{n^4\pi^4}\right)\right]^{1/2}$$

The contribution of the $O(1/n^4\pi^4)$ terms to the sum for $n\geqslant 5$ is in the fifth significant place, and so can be neglected in the evaluation of terms in the summation beyond the second. (This holds at room temperature and some way below. At low temperatures the convergence is poor, as stated earlier, and the second method has to be used.) At higher n, the $\sqrt{2}[\ ]^{1/2}$ term further reduces to

$$(\eta_C+\eta_V)\left[1-\frac{(\eta_C-\eta_V)^2}{8n^2\pi^2}\right] \tag{6}$$

Now

$$\eta_C+\eta_V = 2\eta-x_V-x_C$$

and

$$x_V-\xi = \eta-x_C;$$

by subtraction

$$\eta_C+\eta_V = \eta-\xi$$

also

$$\eta_C-\eta_V = x_V-x_C$$

so that the total $\{\ \}_n$ term in (5a) becomes

$$(\eta-\xi)\frac{(x_V-x_C)^2}{8n^2\pi^2}$$

(It can easily be shown that this approximation for the higher terms holds also for the non-symmetrical case, i.e. when $x_V-\xi\neq\eta-x_C$.)

For $m\geqslant 5$, the remainder from the mth term onwards is then easily approximated by

$$R_m = \frac{(\eta-\xi)(x_V-x_C)^2}{8\pi^2}\sum_{n>m,\,\text{odd}}\frac{1}{n^2}$$

$$= \frac{(\eta-\xi)(x_V-x_C)^2}{8\pi^2}\left(\sum_{n>0,\,\text{odd}}^{\infty}\frac{1}{n^2}-\sum_{n>0,\,\text{odd}}^{m}\frac{1}{n^2}\right)$$

$$= \frac{(\eta-\xi)(x_V-x_C)^2}{8\pi^2}\left(\frac{\pi^2}{8}-\sum_{n>0,\,\text{odd}}^{m}\frac{1}{n^2}\right)$$

Neglecting terms $O(1/n^4\pi^4)$ for *all* n enables one to write a very simple approximate equation for the tunnel characteristic, since then

$$\sum_{n>0,\,\text{odd}}\{\ \}_n = \frac{(\eta-\xi)(x_V-x_C)^2}{8\pi^2}\sum_{n>0\ \text{odd}}\frac{1}{n^2}$$

$$= \frac{(\eta-\xi)(x_V-x_C)^2}{64} \tag{7}$$

Noting that $\eta-\xi$ is the applied voltage in units of kT/q, while x_V-x_C is the corresponding overlap of the band edges in units of kT, the expression for I_T becomes

$$I_T = \frac{A'}{T}V(V_0-V)^2 \tag{8}*$$

where A' = temperature-independent constant and $V_0 = (E_V-E_C)/q$ at equilibrium.

It should be pointed out that, in this approximation, the expression remains the same in the non-symmetrical situation. The difference between the two cases must thus lie in the leading terms of the respective series. The I_T vs. V curve is shown in Fig. 3 for $V_0 = 0{\cdot}2$ V.

LOW-TEMPERATURE EXPANSION

We shall now derive an asymptotic series for I which is useful at low temperatures. Supposing first that $E_C<E_F<E_V$, we write

$$I = \int_{E_C}^{E_F}(E_V-E)^{1/2}(E-E_C)^{1/2}dE$$

$$-\int_{E_C}^{E_F}\frac{(E_V-E)^{1/2}(E-E_C)^{1/2}}{\exp[\beta(E_F-E)]+1}dE$$

$$+\int_{E_F}^{E_V}\frac{(E_V-E)^{1/2}(E-E_C)^{1/2}}{\exp[\beta(E-E_F)]+1}dE \tag{9}$$

Abbreviating the first term by B (which is an elementary integral and independent of tempera-

* Since this approximate expression is valid only in the room-temperature range, and in any case the temperature variation of the tunnel current depends on a number of factors not considered in the simple Esaki theory, little significance can be attached to the predicted T^{-1} dependence.

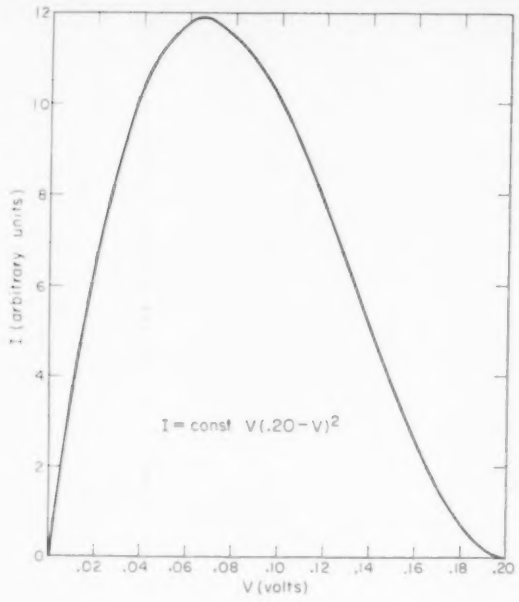


FIG. 3.

ture), we define a characteristic energy ϵ to be the smaller of the two numbers $(E_V-E_F)/2$ and $(E_F-E_C)/2$. We shall now suppose $\beta\epsilon$ to be large. Truncating the last two integrals at $E_F\pm\epsilon$, we find

$$I = B-\int_{E_F-\epsilon}^{E_F}\frac{(E_V-E)^{1/2}(E-E_C)^{1/2}}{\exp[\beta(E_F-E)]+1}\,dE+$$

$$+\int_{E_F}^{E_F+\epsilon}\frac{(E_V-E)^{1/2}(E-E_C)^{1/2}}{\exp[\beta(E-E_F)]+1}\,dE+R_1(\beta) \tag{10}$$

A more careful analysis shows that

$$|R_1(\beta)|<\tfrac{1}{2}(E_V-E_C)^2\,e^{-\beta\epsilon}$$

Using Taylor's series with a remainder, for

$$|E-E_F|<\epsilon \text{ and any } M,$$

$$(E_V-E)^{1/2}(E-E_C)^{1/2}=$$

$$=\sum_{m=0}^{M}a_m(E_F)(E-E_F)^m+a_{M+1}(E_F+\theta\epsilon)\times$$

$$\times\,(E-E_F)^{M+1}$$

Here $|\theta|>1$, and the $a_m(E_F)$ are obtained by repeated differentiation at E_F. Inserting this, we find

$$I = B-\sum_{m=0}^{M}(-1)^m a_m\int_{E_F-\epsilon}^{E_F}\frac{(E_F-E)^m\,dE}{\exp[\beta(E_F-E)]+1}$$

$$+\sum_{m=0}^{M}a_m\int_{E_F}^{E_F+\epsilon}\frac{(E-E_F)^m\,dE}{\exp[\beta(E-E_F)]+1}+R_1(\beta)+R_2(\beta) \tag{11}$$

and again we have an estimate for $R_2(\beta)$:

$$|R_2(\beta)|<\frac{4(E_V-E_C)^2M!}{\pi(\beta\epsilon)^{M+2}}$$

Thus

$$I = B+\sum_{m=0}^{M}[1+(-1)^{m+1}]\frac{a_m}{\beta^{m+1}}\times$$

$$\times\int_0^{\beta\epsilon}\frac{x^m\,dx}{e^x+1}+R_1(\beta)+R_2(\beta)$$

Pushing the upper limit of integration to infinity and collecting the remainder, we have, upon noting that only terms with odd m appear,

$$I = B+2\sum_{k=1}^{K}\frac{a_{2k-1}}{\beta^{2k}}\int_0^{\infty}\frac{x^{2k-1}dx}{e^x+1}+R_1(\beta)$$

$$+R_2(\beta)+R_3(\beta) \tag{12}$$

$$|R_3(\beta)|<\frac{3(E_V-E_C)^2(M+1)!}{2(\beta\epsilon)}\,e^{-\beta\epsilon}$$

Finally, we write

$$I = B+2\sum_{k=1}^{K}\frac{(2k-1)!\,c_k a_{2k-1}}{\beta^{2k}}+r_K(\beta) \tag{13}$$

with

$$c_k=\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^{2k}}$$

Vol. 2
1961

and

$$|r_K(\beta)| < [\tfrac{1}{2}e^{-\beta\epsilon} + \frac{4(2K-1)!}{\pi(\beta\epsilon)^{2K+1}} + \frac{3(2K)!}{2(\beta\epsilon)}e^{-\beta\epsilon}] \times \times (E_V - E_C)^2$$

The first term of the (obviously divergent) series is computed from $c_1 = \pi^2/12$ and

$$a_1(E_F) = \frac{(E_V+E_C)/2 - E_F}{(E_V-E_F)^{1/2}(E_F-E_C)^{1/2}}$$

[Likewise, all $a_{2k-1}(E_F) = 0$ when $E_F = (E_V+E_C)/2$, in which case $B = \pi(E_V-E_C)^2/16$].

As an illustration of the use of the remainder, we take $K = 1$ and $\beta\epsilon = 5$. In this case

$$|r_1(5)| < (0{\cdot}018)(E_V - E_C)^2$$

In general, when $\beta\epsilon$ is given, we should choose that K which minimizes the estimate of the remainder.

For the cases where $E_F < E_C$ or $E_F > E_V$, the work is simpler. First assume $E_F < E_C$. Then

$$I = \int_{E_C}^{E_V} \frac{(E_V-E)^{1/2}(E-E_C)^{1/2}}{\exp[\beta(E-E_F)]+1}\,dE = r'(\beta) \quad (14)$$

say. We have

$$|r'(\beta)| < \frac{\pi}{8}(E_V-E_C)^2 \exp[-\beta(E_C-E_F)]$$

If $E_F > E_V$,

$$I = \int_{E_C}^{E_V} (E_V-E)^{1/2}(E-E_C)^{1/2}dE - \int_{E_C}^{E_V} \frac{(E_V-E)^{1/2}(E-E_C)^{1/2}}{\exp[\beta(E_F-E)]+1}\,dE$$

$$= \frac{\pi}{8}(E_V-E_C)^2 + r''(\beta) \quad (15)$$

where

$$|r''(\beta)| < \frac{\pi}{8}(E_V-E_C)^2 \exp[-\beta(E_F-E_V)]$$

The asymptotic form of I thus depends strongly upon the position of E_F with respect to the interval (E_C, E_V). Only when $E_C < E_F < E_V$ does the series make its appearance, complemented by a large leading term and a small remainder. This change of form of I as E_F varies shows that E_C and E_V are "turning points" of the asymptotic expansion. One effect of the change of form as E_F crosses a turning point is to lessen the practical usefulness of the asymptotic expression near that point. For, with β fixed, as E_F tends to E_C or E_V from within the interval, $\beta\epsilon$ tends to zero and the remainder $r_K(\beta)$ is no longer small. Similar difficulties appear as E_F tends to E_C or E_V from without the interval. Thus, for purposes of practical computation over the range of interest, it will be necessary to exclude the regions where $|E_F-E_C|$ or $|E_V-E_F|$ is not $\gg 1/\beta$. If values of the integral are required across these regions, some form of interpolation between values outside the regions may be used.

Note added in proof: Somewhat similar calculations have recently been published by C. W. BATES, *Phys. Rev.* **121**, 1070 (1961).

REFERENCES

1. L. ESAKI, *Phys. Rev.* **109**, 603 (1958).
2. T. P. BRODY, *J. Appl. Phys.* To be published.
3. T. P. BRODY, *Westinghouse Research Memo.* 406FF451-1-M1 (1960).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 216–221. Printed in Great Britain

MOLTEN–DIFFUSED SILICON TRANSISTORS

SHOJI TAUCHI

Central Research Laboratory, Hitachi Ltd., Kokubunji, Tokyo

(*Received* 12 *August* 1960; *in revised form* 4 *December* 1960)

Abstract—While the author was engaged in the study of producing single crystals for high-frequency transistors based on the conventional grown–diffused process, it was found highly difficult to produce by this method single-crystal material with uniform base width less than a few microns. The disadvantage of the conventional grown–diffused process could be eliminated by using a new method in which impurities are diffused from the molten semiconductor directly into the single-crystal bulk. This single-crystal-producing method was given the name "molten–diffused process". With the application of this process, it has become possible to obtain silicon transistors with a base width less than 1 μ and a cut-off frequency of 580 Mc/s. This paper reports on the transistors so produced.

Résumé—Pendant que l'auteur était engagé dans l'étude de production de cristaux simples pour les transistors à haute fréquence basée sur le procédé diffusé–développé conventionnel, il a trouvé très difficile de produire par cette méthode un matériau à cristal simple ayant une largeur de base uniforme plus petite que quelques microns. Le désavantage du procédé diffusé–développé conventionnel peut être éliminé en utilisant une nouvelle méthode dans laquelle les impuretés sont diffusées, du semiconducteur fondu directement dans le volume du cristal simple. Cette méthode de production du cristal simple a été nommée "procédé Fonte–Diffusion". En applicant ce procédé, il a été possible d'obtenir des transistors au silicium ayant une largeur de base de moins de 1 μ et une fréquence d'interruption de 580 Mc/s. Cet article décrit aussi les transistors produits par cette méthode.

Zusammenfassung—Bei der Herstellung von Einkristallen für Hochfrequenztransistoren nach dem konventionellen Wachstum–Diffusionsverfahren hatten wir grosse Schwierigkeiten bei der Erzeugung eines Einkristallkörpers mit einer einheitlichen Basisdicke von weniger als einigen Mikron. Diese Schwierigkeit liess sich durch ein neues Verfahren umgehen, bei dem die Fremdatome aus dem geschmolzenen Halbleiter direkt in den Einkristallkörper eindiffundiert wurden. Wir nannten dies das Schmelz–Diffusionsverfahren. Durch seine Anwendung erhielten wir Siliziumtransistoren mit einer Basisdicke von weniger als 1 μ und einer Grenzfrequenz von 580 MHz. Wir berichten im vorliegenden Artikel über die auf diese Weise hergestellten Transistoren.

Vol. 2
1961

1. INTRODUCTION

IN THE so-called grown–diffused process,[1] single-crystal material for transistors, silicon for instance, is produced as follows. An *n*-type single crystal with relatively high resistivity (about 1 Ω-cm) is grown from molten silicon doped with antimony at a pulling speed of 1 mm/min by the pulling method. During the course of this growth, the pulling is stopped and the emitter impurities (such as Ga and As) are added in proper amounts into the molten silicon so as to grow an *n*-type emitter section with relatively low resistivity. Owing to the inter-solid phase impurity diffusion during the pulling of the single crystal, gallium, with higher diffusion speed, diffuses from the emitter side into the collector side. At this portion, therefore, a thin *p*-type layer is formed, resulting in an *n–p–n* junction. The author's experience in the grown–diffused process indicates that an attempt to shorten the diffusion time so as to obtain *p*-type base layers as thin as possible is accompanied by the local non-uniformity in the thickness of the *p*-type layer and partial areas where *n–p–n* junctions are not formed, as shown in Fig. 1. The reason for the above is explained as follows. The variation with time in the impurity concentration

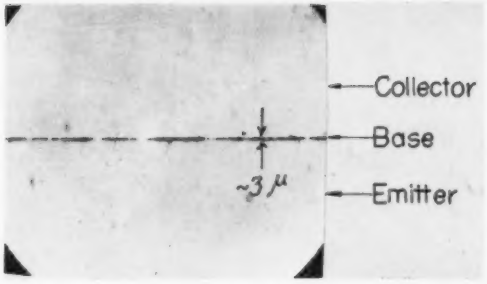


FIG. 1. An *n–p–n*-junction of a grown-diffused Si single crystal.

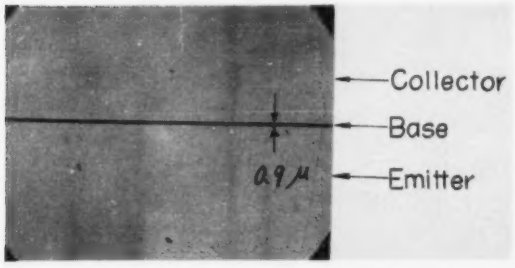


FIG. 2. An *n–p–n*-junction of a molten-diffused Si single crystal.

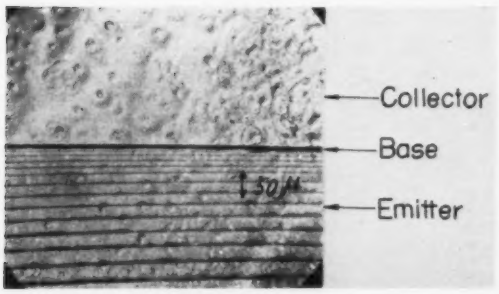


FIG. 3. The striation pattern in the emitter region of a molten-diffused Si single crystal.

facing p. 216

Vol. 2
1961

at the growing surface of the single crystal after the impurity had been added in the emitter region is different according to the distance from the axis of the crystal. This difference results in a local variation of the impurity distribution in the crystal; hence in the non-uniformity in the width of the *p*-type base region, and, in some cases, to the occurrence of local areas shown in Fig. 1, where an *n*–*p*–*n* junction is not formed. For this reason it is difficult to produce *n*–*p*–*n*-junction single crystals with extremely thin and uniform base width.

2. MOLTEN–DIFFUSED PROCESS

In this section, a description of the molten–diffused process is given. Molten silicon, with proper amounts of donor and acceptor impurities admixed uniformly, is prepared. A previously obtained collector single crystal is slowly and slightly contacted to this molten silicon, thereafter waiting for conditions of thermal equilibrium to be reached between the solid and liquid states of silicon. It takes several to tens of minutes for the condition of thermal equilibrium to be established in this case. As the melting speed of solid-phase silicon decreases, mainly the acceptor impurity diffuses from the melt into the single crystal, thereby forming a *p*-type layer. This layer varies possibly from 0 to 10 μ in thickness, and, therefore, it is difficult to obtain a thin, uniform *p*-type layer with some 1 μ thickness. To eliminate this difficulty, an attempt was made: the temperature of the molten silicon is slightly elevated so that the single crystal melts slowly and uniformly until it reaches the state in which both melting and growing phenomena completely stop; while maintaining this state for a required period the acceptor impurity with high diffusion speed is allowed to diffuse from the molten silicon directly into the single crystal. In this case, the speed at which the single crystal melts must be lower than the diffusion speed of impurities within the melt but higher than that of impurities in the solid. To meet this requirement, it is necessary, of course, that the temperature of the crucible should be controlled to within $\pm 0{\cdot}1$°C and temperature of the molten silicon be varied in accordance with a predetermined schedule. After this, the temperature of the molten silicon is quickly lowered so as to cause rapid growth of *n*-type single crystals for the emitter region. This process offers the *n*–*p*–*n*-junction single-crystal material having an extremely thin and uniform base layer, as shown in Fig. 2.

The reason is as follows. Firstly, if the doping level of donor impurity in the collector region is suitably selected, the concentration distribution curves of diffused donor and acceptor impurities intersect each other, resulting in the formation of a *p*-type base layer however short the diffusion time may be. Secondly, because the melt of silicon with desired impurities admixed uniformly is previously prepared, there is no non-uniformity caused in impurity-concentration distribution depending upon the distance from the crystal axis, as is seen in the grown–diffused process. Thus, the width of the base region formed by the diffusion can be made highly uniform. Thirdly, since the solid–liquid interface is an isothermal surface when the crystal stops both melting and growing, the temperature of the diffusion source is constant; and this makes it easier to obtain an extremely uniform and sharp diffused layer. Fourthly, since the solid silicon is kept at an elevated temperature just below its melting point, the impurity diffusion into it is hardly affected by the lattice defects such as dislocation or anisotropy. Finally, only the required part of the collector single crystal can be cut away, the remainder of the single crystal being put into repeated use.

Fig. 3 shows an *n*–*p*–*n* junction obtained in the way mentioned above and treated with relatively severe staining. A thin, long band in the centre represents the base region.

To the above and underneath of this band there are collector and emitter regions, respectively. Striations appearing in the emitter region have intervals which are quite narrow in the vicinity of the junction and gradually become wider with increase in their distance from the base. This indicates well that in the molten–diffused process both melting and growth first stop along the {111} crystallographic plane perpendicular to the growth axis and then the melt is rapidly cooled, so, as a result, the growth speed of the crystal gradually increases. If it is assumed in this case that the severe supercooling in the course of the growth of the single crystal causes the growth to be made intermittently, then the formation of the striations can be explained. It is considered that the existence

of such a supercooling phenomenon in the molten–diffused process makes it relatively easy to realize the maintenance of a single crystal in a state where both the growth and melting stop for a required time. The etched striation pattern caused by the intermittent growth of crystal is already observed on the single-crystal germanium made by the zone-levelling method.[2]

3. THICKNESS OF THE BASE REGION

If the diffusion time is short, the base and emitter impurities diffuse mainly out of the melt under the condition where the crystal is neither growing nor melting, and slightly within the solid phase during the subsequent growth of crystal of the emitter region. If the time of diffusion from the melt and the pulling time of the emitter region are denoted by t_m and t_s, respectively, then the overall diffusion time t_d of the impurity becomes

$$t_d = t_m + t_s \qquad (1)$$

During the process in which melting of the collector single crystal is stopped, the impurity concentration of the diffusion source, in other words, the impurity concentration under the solid–liquid interface of the collector crystal, is a little lower than that in the melt.

In the course of the melting process of the collector single crystal, most of the melted silicon is mixed with the molten silicon originally present in the crucible, but a thin layer of the melted silicon remains in contact with the newly appeared crystal surface.

Therefore, the base and emitter impurities are not contained in this layer at first. These impurities diffuse through this thin layer and arrive at the solid surface.

If the speed of melting of the collector crystal is held constant, effective thickness δ of this layer can be determined from the lower threshold of the impurity concentration in the melt to obtain the base layer, for a certain impurity concentration of the collector crystal. By this method we determined $\delta = 0{\cdot}16$ mm when the melting speed of the collector crystal was about 20 μ/sec. The value of δ is also determined independently by comparing the calculated thickness of the base layer with the observed value. This value was about 0·1 mm. The difference between the two is probably due to the experimental error in determining the base-layer thickness.

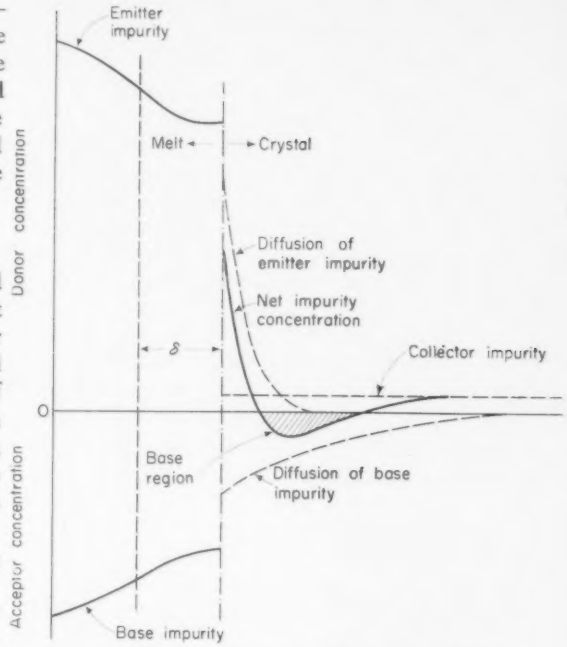


FIG. 4. Impurity distribution in the solid- and liquid-state silicon in the molten–diffused process.

The impurity concentration in this process is shown in Fig. 4.

When the impurity diffuses out of the melt during t_m, impurity concentration C within the solid is given approximately as

$$C = \frac{kC_0}{1+\mu k}\operatorname{erfc}\left(\frac{x+\mu\delta}{2\sqrt{(D_s t_m)}}\right) \simeq kC_0 \exp(-\alpha u^2 - \beta u) \qquad (2)$$

where

$$u = \frac{x+\mu\delta}{2\sqrt{(D_s t_m)}}, \quad \mu = \sqrt{(D_s/D_m)} \ll 1$$

The coefficients α and β in equation (2) are obtained by the method of least squares under the experimental condition

$$[x/2\sqrt{(D t_m)} = 0 \sim 3]$$

Vol. 2
1961

as follows

$$\alpha = 0{\cdot}927 \qquad \beta = 0{\cdot}824$$

C_0 is the impurity concentration in the molten silicon, k is the distribution coefficient of the impurity for silicon, and D_s and D_m are the diffusion coefficients of the impurity in solid and liquid silicon respectively.

With the initial conditions given by equation (2), the impurity concentration distribution after the diffusion for t_s sec at inter-solid phase is given as

$$C(x, t_s) = \frac{kC_0}{2}\operatorname{erfc}\xi_0 \cdot \operatorname{erfc}\xi + \frac{kC_0}{2}\frac{1}{\sqrt{(1+\alpha\eta)}}$$

$$\times \exp\left\{-(\xi^2+\alpha\xi_0^2+\beta\xi_0)+\left(\frac{\xi-\frac{\beta}{2}\sqrt{(\eta)}-\alpha\xi_0\sqrt{(\eta)}}{\sqrt{(1+\alpha\eta)}}\right)^2\right\}$$

$$\times\left\{1+\operatorname{erf}\left(\frac{\xi-\frac{\beta}{2}\sqrt{(\eta)}-\alpha\xi_0\sqrt{(\eta)}}{\sqrt{(1+\alpha\eta)}}\right)\right\} \quad (3)$$

where

$$\xi = x/2\sqrt{(D_s t_s)}, \quad \xi_0 = \delta/2\sqrt{(D_m t_m)}, \quad \eta = t_s/t_m$$

As for the effect of change in temperature in the vicinity of the base due to the pulling of the single crystal, the effective diffusion time t_{eff} is considered instead of t_s and defined as

$$t_{\text{eff}} = \frac{1}{D(T_m)}\int_0^{t_s} D(T)\,dt$$

$$\simeq \{1-\exp(-t_s/\tilde{t})\}\,\tilde{t} \quad (4)$$

where

$$\tilde{t} = (RT_m^2/\Delta H) f \cdot \frac{dT}{dz}$$

and $D(T_m)$ and $D(T)$ denote the diffusion coefficients at the melting point T_m of silicon and at an arbitrary temperature T, respectively. ΔH is the activation energy of diffusion in the solid and R is the gas constant; dT/dz is the temperature gradient in the crystal along the growth direction and f is the pulling rate of the emitter region. For silicon, values of $D(T)$ and ΔH have already been obtained.[3] From equation (3), we can obtain the distance from solid–liquid interface to the collector-base and emitter-base junction respectively. So the relationship between the base width W_b and the diffusion time $t_d = t_m + t_s$ is obtained as shown in Fig. 5. For these calculations gallium and arsenic were assumed as the base and the emitter impurities, respectively, with both of their concentrations in the molten silicon $2\times10^{19}/\text{cm}^3$. The donor impurity concentration in the collector region was taken to be $5\times10^{15}/\text{cm}^3$ and the value of t_m was kept constant at 10 sec while the value of t_s was varied. Measured values are also shown in the same figure, and their scatter about the curve is considered to be due to error in the measured values of the base width. From the figure, it can be seen that a base width less than 1 μ requires the diffusion time t_d to be limited to a value less than about 20 sec.

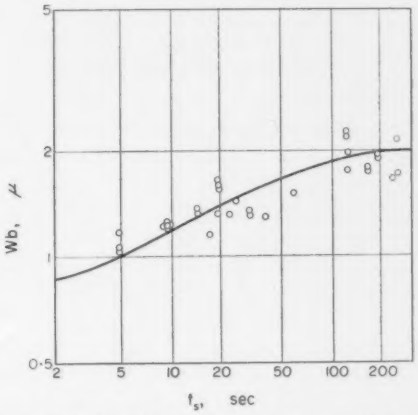


FIG. 5. Relationship between the base width and the diffusion time (t_s) in the molten–diffused process.

In the case of the grown–diffused process, if we assume that the impurity concentration distribution before the diffusion takes the ideal form of a step function and the impurities diffuse only in the course of growth of the emitter region, then the distribution after the diffusion is given by

$$C = \frac{kC_0}{2}\operatorname{erfc}\left(\frac{x}{2\sqrt{(D_s t_{\text{eff}})}}\right) \quad (5)$$

The values of the base width, W_b^{cal}, with respect to the various diffusion times, were calculated

from equation (5). The ratio of these values to the measured values of base width, W_b^{bs}, for the twenty-five crystals with different diffusion times was found to be in the range of $0{\cdot}77 \pm 0{\cdot}23$, which indicates that the calculated values are considerably smaller than the observed ones.

On the other hand, if we take into account all the phenomena involved in the molten–diffused process, such as the existence of a thin layer of melted silicon, impurity diffusion from melt into solid, and diffusion through solid in the course of the growth of the emitter region, the ratio of W_b^{cal} obtained by equation (3) to the corresponding observed values was in the range of $1{\cdot}01 \pm 0{\cdot}15$, showing agreement between the calculated and the observed values within the range of the experimental error. It became clear from these results that, in the case of the molten–diffused process, the values of the base width obtained by calculation, taking into account only the inter-solid phase impurity diffusion during the pulling of the emitter region, were considerably lower than the observed ones.

4. CHARACTERISTICS OF MOLTEN–DIFFUSED TRANSISTORS

This section describes the characteristics of transistors which were constructed from the silicon n–p–n-junction single-crystal material obtained by the process described so far. Because of uniformity in diffusion around the axis, the single crystal in this case has uniform base thickness over the entire crystal, and hence has quite uniform electrical characteristics. So it is easy to produce hundreds of transistors with cut-off frequency higher than 300 Mc/s from the same crystal at high yield. Table 1 lists the characteristics of such transistors. The collector-junction capacitance is appreciably small. In addition, the base-to-emitter junction capacitance of this transistor is relatively small. These facts may be ascribed to the thin intrinsic semiconductor regions which exist on those junctions. There are various factors determining the cut-off frequency of such transistors.[4] But, in general, it may be given as

$$\frac{1}{f_\alpha} = \frac{1}{f_B} + \frac{1}{f_E} + \frac{1}{f_C} \tag{6}$$

The first, second and third terms on the right-hand side of equation (6) correspond to time constants of base region, emitter junction and collector junction, respectively; f_E and f_C are inversely proportional to the product of emitter-junction capacitance and emitter resistance and the product of collector resistance and collector-junction capacitance, respectively. Details of them will be discussed in a separate article, and, for the time being, it is assumed that the cut-off frequency f_α is determined mainly by the cut-off frequency of the base region, f_B. Then, if the effect of the drift field within the base is disregarded, we have

$$f_\alpha = \frac{D_n}{\pi W_b^2} \tag{7}$$

where W_b denotes the base width (cm) and D_n the diffusion coefficient of minority carriers within the base. For a silicon n–p–n junction, D_n is about 20 cm^2/sec,[5] if resistivity of the base is assumed to be $0{\cdot}5\ \Omega$-cm. Since the base width of the crystal with diffusion time less than 20 sec is found to be about $1\,\mu$ from Fig. 5, the cut-off frequency given by equation (7) is about 600 Mc/s. The value of $f_\alpha = 580$ Mc/s given in the table is in good agreement with this. As for the base-lead resistance

Table 1. *Static characteristics of molten–diffused silicon transistors* ($T_a = 25°C$)

Item	Measuring conditions	Typical characteristics
α-Cut-off frequency (f_α)	$V_C = 6$ V, $I_E = -2$ mA	580 Mc/s
Collector capacitance (C_c)	$V_C = 6$ V, $I_E = 0$	1·0 pF
Emitter-junction capacitance (C_{te})	$I_E = 0$	1·9 pF
Input impedance (h_{te})	$V_C = 6$V, $I_E = -2$ mA, 100 Mc/s	$56 - j\,182\ \Omega$

Vol. 2 1961

$r_{bb'}$, it cannot be treated as one pure resistance in the case where the base lead of this type of transistor is connected to only one side of the junction bar, but, because of the internal-carrier flow lines being collected to the base-lead side, a condenser should be added in parallel with the resistance in the equivalent circuit.[6] These two parameters can be obtained by analysing the results of the measurement of input impedance in the emitter-grounded-type circuit over a wide frequency range. The input impedance h_{ie} in the table is considered to have rather a large imaginary part, but, if only the real part is taken into consideration, it is not necessarily too large. It is expected that the base-lead resistance of this type of transistor can be made as small as that of the general mesa-type or drift-type transistors. Moreover, it is found that additional electrostatic capacitance due to the mounting of the base lead is introduced to emitter-to-base and collector-to-base as stray capacitance, and this affects each parameter.[7] With regard to this, detailed study will be continued in future.

Fig. 6 shows an example of the frequency behaviour of the current amplification factor α with the grounded base of this transistor measured with a Transfer Function and Immittance Bridge. As seen from the figure, the cut-off frequency at the collector voltage of 6 V and emitter current of -2 mA is about 580 Mc/s. On the lower-frequency side, the absolute value becomes larger than that expected; this characteristic is interpreted to be due to the above-mentioned electrostatic capacitance caused by the mounting of the electrode between the emitter and base.

5. CONCLUSION

The production of n–p–n silicon single-crystals, with base width less than 1 μ, based on the molten–diffused process newly developed by the author, has been discussed. This process made it easy to obtain silicon transistors with cut-off frequency more than 580 Mc/s. Further, the impurity-concentration distribution within the junction crystal

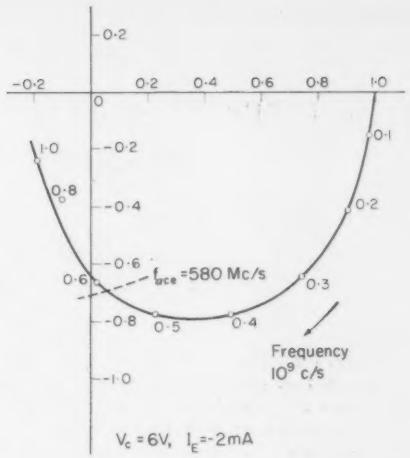


FIG. 6. Frequency behaviour of the current amplification factor of the molten–diffused Si transistor with grounded base.

realized by the same process was analysed. This, together with the examination of the electrical characteristics of the transistor, offered a fairly detailed explanation pertaining to the relationship between the impurity-concentration distribution and the electrical characteristics of the transistor.

Acknowledgements—The author wishes to acknowledge the direction and encouragement of M. TOMONO, the head of the 2nd Department of this laboratory, and also to express his appreciation to K. SATO and T. TOKUYAMA who supplied the data.

REFERENCES

1. B. CORNELISON and W. A. ADCOCK, *IRE Wescon. Conv. Rec.* Pt-3, 22 (1957).
2. H. UEDA, *J. Phys. Soc. Japan.* **16,** 61 (1961).
3. C. S. FULLER and J. H. DITZENSBERGER, *J. Appl. Phys.* **27,** 544 (1956).
4. F. J. HYDE, *Proc. Instr. Elect. Eng. Soc.* **106**b, 397 (1959)
5. D. C. CRONEMEYER, *Phys. Rev.* **105,** 522 (1957); E. M. CONWELL, *Proc. I.R.E.* **46,** 1281 (1958).
6. R. L. PRITCHARD, *Proc. I.R.E.* **46,** 46 (1958).
7. R. L. PRITCHARD, *Proc. I.R.E.* **43,** 38 (1955).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 222–231. Printed in Great Britain

THE POTENTIALITIES OF SILICON AND GALLIUM ARSENIDE SOLAR BATTERIES

T. S. MOSS

Royal Aircraft Establishment, Farnborough, Hants.

(*Received* 21 *October* 1960; *in revised form* 13 *December* 1960)

Abstract—The theory of the spectral response of a *p–n* junction solar battery unit is given, and detailed comparisons are made of the expected performance of silicon and gallium arsenide units. Data now available on absorption in GaAs show that it is a particularly favourable material because it has a very steep absorption edge. The low effective masses and correspondingly low densities of states in GaAs are also advantageous.

The efficiency of practical Si units is unlikely to exceed 15 per cent even with intensive development, whereas GaAs units with efficiencies of about 20 per cent should be achievable in the near future and the performance may ultimately approach 25 per cent, even if (as assumed) the carrier lifetimes stay well below those for Si.

Résumé—La théorie de la réponse spectrale d'un élément de batterie solaire à jonction *p–n* est demontrée, et des comparaisons détaillées du rendement attendu des éléments de silicium et d'arséniure de gallium sont faites. Les données de l'absorption dans l'AsGa qui sont maintenant disponibles démontrent que c'est un matériau particulièrement favorable parce qu'il comporte une arête d'absorption très raide. Les faibles masses effectives et les faibles densités de condition correspondantes dans l'AsGa sont aussi avantageuses.

Le rendement des éléments pratiques de Si ne dépassera probablement pas 15 pour cent même avec un développement intensif tandis que les éléments d'AsGa ayant des rendements d'environ 20 pour cent seraient réalisables très prochainement et le rendement pourrait même atteindre 25 pour cent, même si (comme il est supposé) les durées de vie du porteur seraient plus courtes que celles du Si.

Zusammenfassung—Nach Entwicklung der Theorie der spektralen Empfindlichkeit einer Sonnenbatterie mit *p–n*-Übergang, werden die zu erwartenden Leistungen von Silizium- und Galliumarsenid-Elementen im einzelnen verglichen. Die jetzt verfügbaren Angaben über die Absorption von GaAs zeigen, dass es ein besonders geeignetes Material ist, da es eine sehr steile Absorptionskante besitzt. Die niedrigen Effektivmassen und die entsprechend niedrige Besetzung der Energiezustände in GaAs sind ebenfalls von Vorteil.

Die Leistung der praktischen Siliziumzellen wird auch bei intensiver Entwicklung im Bereich von 15 Prozent bleiben, während sich bei den GaAs-Zellen bald Leistungen bis zu 20 Prozent erzielen lassen sollten. Die endgültige Leistung könnte sogar bis zu 25 Prozent gesteigert werden, auch wenn, wie angenommen wird, die Lebensdauer der Träger beträchtlich unterhalb der von Silizium liegt.

Vol. 2
1961

1. INTRODUCTION

THE first publication describing the use of large-area *p–n* junctions as solar-power converters was by CHAPIN *et al.*[1] These workers used silicon, and nearly all the development work carried out since that time has been on this same semiconductor.

It has become increasingly evident over the past year or two that the efficiency of silicon units is approaching a limit, although the performance obtained (with the latest cells, 13 per cent for appreciable numbers and 14 per cent for a few selected units) is only about half that estimated in early theoretical work.

It is, therefore, timely to investigate in some detail the theoretical behaviour of Si solar batteries in order to find the fundamental causes of this rather disappointing performance, and at the same time apply the theory to gallium arsenide—for

which we have recently carried out the necessary measurements of absorption behaviour (Moss and Hawkins[2]).

Consideration is restricted to single-component systems. The possible performance of some "multi-layer" cells has been discussed by Wolf[3].

2. NOTATION

y, t, d	Thicknesses	(μ)
J^-, J^+	Electron and hole currents	(Unit charge flow)
J_t^-, J_t^+	Electron and hole currents at the junction	(Unit charge flow)
J_s^-, J_{sc}^+, J_{sc}	Electron and hole short-circuit currents	(Unit charge flow)
J_0	Integrated short-circuit current	(Unit charge flow)
j_0	Junction parameter	(Unit charge flow)
n, n_0, n_d, n_i	Electron concentrations	(cm^{-3})
p, p_0, p_a	Hole concentrations	(cm^{-3})
$\Delta n, \Delta p$	Concentrations of photo-carriers	(cm^{-3})
N, P	Functions of carrier concentration	
A, B, R, S	Constants of integration	
τ_e, τ_h	Lifetimes of minority carriers	(sec)
D_e, D_h	Diffusion constants of minority carriers	($cm^2\ sec^{-1}$)
L, l	Diffusion lengths of minority carriers	(μ)
K	Absorption coefficient	(cm^{-1})
X	$\exp(-Kt)$	
F, G	Functions of K, L and l	
s, α, α'	Surface recombination parameters	
I, Q_0	Photon fluxes	($cm^{-2}\ sec^{-1}$)
E	Energy gap	(eV)
k	Boltzman's constant	(eV/°K)
T	Absolute temperature	(°K)
$V, V_{oc},$	Junction voltages	(V)
W, W_{opt}	Power from the junction	(W)
λ	Wavelength	(μ)
r	Junction parameter	
e	Unit charge	(C)
$\underset{\sim}{k}$	Wave vector	
π	Circumference/diameter	
$m_0, m_{eff}, m_e^*, m_h^*$	Carrier masses	
η, η_J	Efficiencies	

3. THEORY OF THE SPECTRAL RESPONSE OF A *p–n* JUNCTION

The configuration treated and the notation used are similar to those of Moss[4] (Chap. 4). As shown diagrammatically in Fig. 1, the junction is assumed

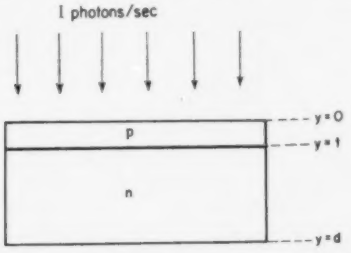


Fig. 1. Diagrammatic representation of solar battery.

to consist of a thin p-layer extending from $y = 0$ to the junction at $y = t$, irradiated in the y-direction at $y = 0$ so that I photons/sec are absorbed per unit area. The n-layer extends from $y = t$ to $y = d$. The behaviour of the minority carriers in the p- and n-regions will be considered in turn. The problem is taken to be one-dimensional (i.e. edge effects are ignored) and, to simplify the notation, currents will be expressed as flows of unit charges.

3.1 *In the p-region, for the electrons*

The continuity equation (corresponding to equation 4.15) is

$$dJ^-/dy = \Delta n/\tau_e - KI\exp(-Ky) \qquad (1)$$

where τ_e is the lifetime of electrons in the p-region.

The current equation (4.13) is replaced by

$$J^- = D_e\, d\Delta n/dy \qquad (2)$$

if we ignore the field component of current compared with the diffusion current. This is a good approximation when the field is due only to the Dember voltage, which is very small in heavily doped material.

The solution of these equations is

$$\Delta n = A\cosh y/L + B\sinh y/L - (FL/D_e)\exp(-Ky) \tag{3}$$

where the diffusion length

$$L = (D_e\tau_e)^{1/2} \text{ and } F = KIL/(K^2L^2-1).$$

The boundary conditions are

$$y = 0, \quad J^- = D_e\, d\Delta n/dy = s\Delta n \tag{4}$$

$$y = t, \quad n_t = n_0 \exp eV/kT \tag{5}$$

where n_0 is the electron concentration in the dark, V the voltage across the junction, and s the surface recombination velocity.

Evaluating A and B we find the current at the junction ($y = t$) to be

$$J_t^- = FKL\exp(-Kt) - F(\alpha+KL)/(\cosh t/L + \alpha\sinh t/L) + [N+F\exp(-Kt)]\tanh(\alpha'+t/L) \tag{6}$$

where $\tanh\alpha' = \alpha = \tau s/L$ and

$$NL/D_e = n_0(\exp eV/kT - 1)$$

This expression is readily shown to be finite at $KL = 1$ where $F \to \infty$.

For short wavelengths where K is large (i.e. $KL \gg 1$) and for typical conditions where $\alpha < t/L < 1$, the short-circuit current ($V = 0 = N$) becomes

$$J_{sc}^-/I = (\cosh t/L + \alpha\sinh t/L)^{-1} - \exp(-Kt) \tag{7}$$

For maximum efficiency it is thus clearly necessary to make α as small as possible, and t/L as small as is compatible with the requirements of absorption in the layer and transverse-layer resistance.

3.2 *In the n-region, for the holes*

The primary equations are similar to (1) and (2) namely

$$dJ^+/dy = KI\exp(-Ky) - \Delta p/\tau_h \tag{8}$$

$$J^+ = -D_h\, d\Delta p/dy \tag{9}$$

giving

$$\Delta p = R\cosh y/l + S\sinh y/l - (Gl/D_h)\exp(-Ky) \tag{10}$$

where

$$l = (D_h\,\tau_h)^{1/2}$$

is the hole diffusion length and

$$G = KIl/(K^2l^2-1) \tag{11}$$

The boundary conditions are

$$\begin{aligned} &y = t, \quad p_t = p_0 \exp eV/kT \\ &y = t+d, \quad \Delta p = 0 \end{aligned} \tag{12}$$

assuming a relatively thick n layer.

Following from the use of a thick layer, we put $\exp(-Kd) = 0$, and $\coth d/l = 1$ and obtain for the current at the junction due to holes

$$J_t^+ = P - G(Kl-1)\exp(-Kt) \tag{13}$$

It is interesting to evaluate this expression for the condition $t \to 0$, i.e. the p-layer is so thin that there can be no absorption in it. The short-circuit current (i.e. $V = 0$ and hence $P = 0$) becomes

$$J_{sc}^+ = IKl/(1+Kl) \tag{14}$$

which shows that the full current of I electrons per second will be obtained if $Kl \gg 1$, i.e. if the absorption takes place well within a diffusion length in the n-region. In the practical solar battery, photoholes produced in the n-region will give an important contribution to the total current, particularly at the longer wavelengths where the absorption is such that the p-layer is relatively transparent ($Kt < 1$) but $Kl > 1$. For typical parameters, i.e. $\alpha = 0{\cdot}1$ and $t/l = t/L = 0{\cdot}3$, the current contribution of the n-layer exceeds that of the p-layer when $KL > 1{\cdot}5$, i.e. for $\lambda > 0{\cdot}75\mu$ for silicon.

3.3 *Current at the junction*

Writing for convenience $\exp(-Kt) = X$, the currents at the junction [equations (6) and (13)] become

$$J_t^- = FKLX - F(\alpha+KL)/(\cosh t/L + \alpha\sinh t/L) + (N+XF)\tanh(\alpha'+t/L) \tag{15}$$

$$J_t^+ = -G(Kl-1) + P = P - KIlX/(1+Kl) \tag{16}$$

Vol. 2
1961

The total short-circuit junction current (i.e. $N = P = 0$) is thus

$$-J_{sc} = F\left\{\frac{\alpha+KL}{\cosh t/L+\alpha\sinh t/L} - KLX - X\tanh(\alpha'+t/L)\right\} + \frac{KlX}{1+Kl} \qquad (17)$$

In discussing this expression, it will be seen that the useful terms are the first and last, the second and third terms represent losses. For reasonably high absorption conditions the second and last terms largely cancel—showing that what radiation is transmitted through the p-layer is used effectively in the n-layer provided that it is absorbed within a diffusion length. Clearly the first term should be maximized and the third minimized. Both trends result from reducing α, so it is clearly advisable to keep the surface recombination small, although the effect is not very marked.

Increase in L is always advantageous as it reduces t/L, whilst increases in either diffusion length extend the waveband over which absorption is effective.

Equation (17) shows only slight dependence on the *relative* magnitudes of l and L. It is not, therefore, automatically preferable to have the higher mobility minority carriers in the surface layer, as is common design practice, but on the contrary it may be advantageous to have a surface n-layer with more mobile majority carriers and lower transverse resistance.

As the relative magnitudes of l and L have little effect, it is convenient to put them equal in equation (17) and obtain

$$-J_{sc}/I = \frac{KL[\alpha+KL-(1+\alpha)\exp(t/L-Kt)]}{(K^2L^2-1)(\cosh t/L+\alpha\sinh t/L)} \qquad (18)$$

This expression, which is readily seen to be finite even when $KL = 1$, is plotted in Fig. 2 as a function of the dimensionless parameters KL and t/L. It has a maximum (for any absorption level such that $KL > 1$) when t is given by

$$\exp Kt = \coth t/L \exp t/L - 1/(KL\sinh t/L) \qquad (19)$$

or approximately by

$$Kt(\exp Kt - 1) = KL - 1 \qquad (20)$$

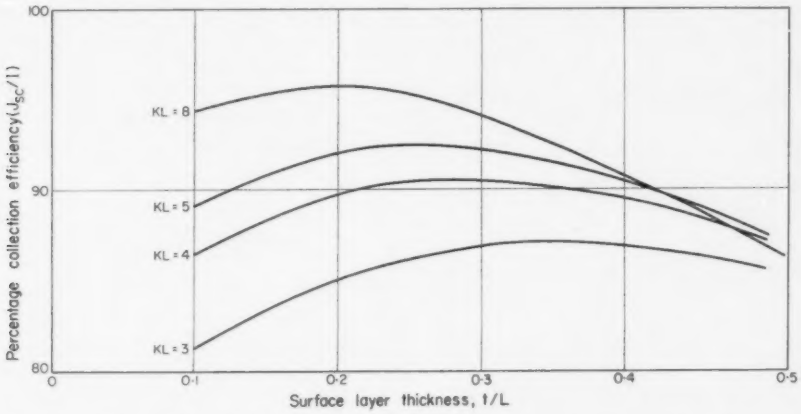


FIG. 2. Calculated collection efficiencies for various absorption levels.

If the useful range of K is thus 2500–10,000 cm^{-1} and $L \sim 8\mu$, the surface layer should be 1–3μ thick. The curves however are very flat, and a value of $t/L = 0{\cdot}25$ will be very near the optimum for $KL = 2$–8. For this value of t/L the "collection efficiency" given by equation (18) will exceed 92 per cent for $KL > 5$. More generally there is also a maximum when $L \neq l$, i.e. for equation (17). In the practical design of a soiar battery it may be advisable to have the surface layer somewhat thicker than this optimum in order to reduce the

transverse resistance through which the output current has to flow.

The total short-circuit current (J_0) will be obtained by integrating equation (18) over all wavelengths, i.e.

$$J_0 = \int_0^\infty -J_{sc}\, d\lambda \text{ electrons/sec} \qquad (21)$$

3.4 *Voltage and power output*

From equations (15) and (16) the general expression for the junction current is

$$J = j_0 (\exp eV/kT - 1) - J_0 \qquad (22)$$

where

$$j_0 = p_0 D_h/l + (n_0 D_e/L) \tanh (\alpha' + t/L) \qquad (23)$$

The open-circuit photo-voltage is thus given by

$$eV_{oc} = kT \log (1 + J_0/j_0) = kT \log J_0/j_0 \qquad (24)$$

as J_0/j_0 is extremely large.

The power delivered by the junction is, from equation (22),

$$VJ = V(j_0 \exp eV/kT - J_0) \qquad (25)$$

This can readily be shown to have its optimum value when

$$V_{opt} = r(kT/e) \text{ and } J_{opt} = -J_0 r/(r+1) \qquad (26)$$

where

$$(r+1)\, e^r = J_0/j_0$$

The optimum power is then

$$W_{opt} = (r-1)\, J_0 kT/e \qquad (27)$$

and the efficiency (at optimum power) is given by

$$\eta = (r-1)\, kT\, J_0/1{\cdot}4 Q_0 \qquad (28)$$

where Q_0 is the total rate of arrival in the semiconductor of solar photons and 1·4 eV is their average energy.

3.5 *Conditions for high efficiency*

Under optimum matching conditions the solar battery is seen to behave as though each photo-carrier reaching the junction provided $(r-1)\ kT$ joules of energy to the external circuit.

Ways of increasing J_{sc} and (hence J_0) in order to have the maximum number of carriers reaching the junction have already been discussed. To increase the energy yield per carrier it is necessary to make r as large as possible, and hence, since $r \sim \log J_0/j_0$, to make j_0 as small as possible. Now

$$j_0 = p_0 D_h/l + (n_0 D_e/L) \tanh (\alpha' + t/L)$$
$$\text{or} \qquad j_0 = p_0 l/\tau_h + n_0 l/\tau_e \qquad (29)$$

as α' should be considerably less than t/L.

Now p_0 and n_0 are the densities of *minority* carriers, given by

$$p_0 = n_i^2/n_d \qquad n_0 = n_i^2/p_a \qquad (30)$$

where n_i is the intrinsic concentration and n_d and p_a are the concentrations of ionized donors and acceptors respectively.

Thus

$$j_0 = n_i^2 \{t/\tau_e p_a + l/\tau_h n_d\} \qquad (31)$$

The main feature of equation (31) is the rapid increase with n_i and thus activation energy. By contrast the other parameters will probably vary little from one semiconductor to another, particularly if the technologies of preparing the semiconductors are at similar states of development. As the efficiency will be seen to depend roughly on the logarithm of j_0, and only fall about 3 per cent for a 3:1 increase in j_0, high accuracy in estimating j_0 is not needed.

In order to reduce j_0 it is thus clearly necessary to use high doping levels. With increasing doping, however, the lifetimes will fall and at high concentrations there will be a tendency for $\tau_h n_d$ and $\tau_e p_a$ to become constant and roughly equal. Also, as it is necessary to maintain long diffusion lengths in order to keep J_0 large, the doping cannot be too high.

It will generally be found that the first term in equation (31) is several times smaller than the second, so that the latter is the important one to minimize. For Si, for example, where typically $p_a = n_d = 2 \times 10^{18}$, $t = 2\mu$, $D_h = 7 \text{cm}^2/\text{sec}$ and $\tau_h n_d \simeq \tau_e p_a = 10^{11}$ cm^{-3} sec we have

$$j_0/n^2 = (2+6)10^{-15} \text{ or } j_0 \sim 10^{-14}\, n_i^2 \qquad (32)$$

Vol. 2
1961

4. DISCUSSION OF THE PROPERTIES OF SILICON AND GALLIUM ARSENIDE

The main compromise in the design of a solar battery is between materials of low energy gap, where most of the solar photons are sufficiently energetic to produce photoelectrons so that J_0 and the short-circuit current are large, and high energy-gap materials where n_i and j_0 are low so that the open-circuit voltage is high. In the simple case where all photons of energy greater than the energy gap are assumed to produce useful photocarriers and all other parameters are taken to be the same for all materials, it may be shown that there is an optimum energy gap, and that it is about 1·4 eV.[4]

4.1 *Absorption characteristics*

It is clearly very important that in a material of given energy gap (and hence given n_i) the absorption should be high for wavelengths right up to those equivalent to the energy gap, i.e. the material should have a steep absorption edge, and in this respect GaAs is markedly superior to silicon. Results of recent measurements of the absorption coefficient for pure single-crystal GaAs, made in this laboratory, are shown in Fig. 3 along with silicon data from DASH and NEWMAN[5] and BRAUNSTEIN *et al.*[6] The edge in GaAs is seen to be very steep, the fundamental reason being that the transitions are vertical as both band extrema are at $\tilde{k} = 0$.

From these curves the spectral distribution of the short-circuit current has been computed and plotted in Fig. 4. On multiplying these curves by the distribution of solar radiation (taking the latter to be that of a black-body at 6000°C) and integrating, we can determine the collection efficiency in terms of short-circuit electron flow per photon, i.e. J_0/Q_0. The results are given in Table 1 for a silicon unit with the usual surface p-layer and for GaAs units with the surface layer of either p- or n-type. It should be noted that no atmospheric

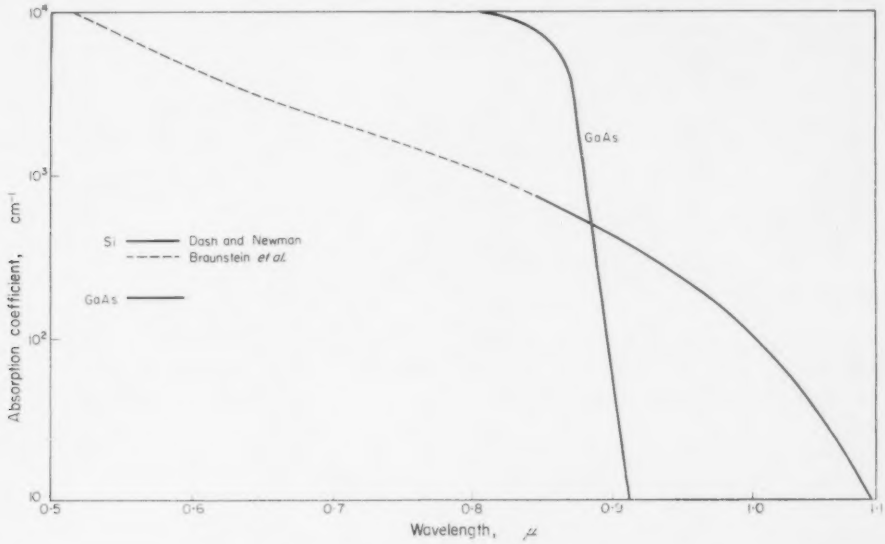


FIG. 3. Absorption in GaAs and Si.

Table 1. *Collection efficiency in electrons per solar photon* (J_0/Q_0)

Dimensions	Silicon	GaAs
(μ)	(%)	(%)
$t = 2, L = l = 8$	37	
$t = 3, L = l = 8$	38	
$t = 2, L = 10, l = 4$		42
$t = 1{\cdot}6, L = 4, l = 10$		39

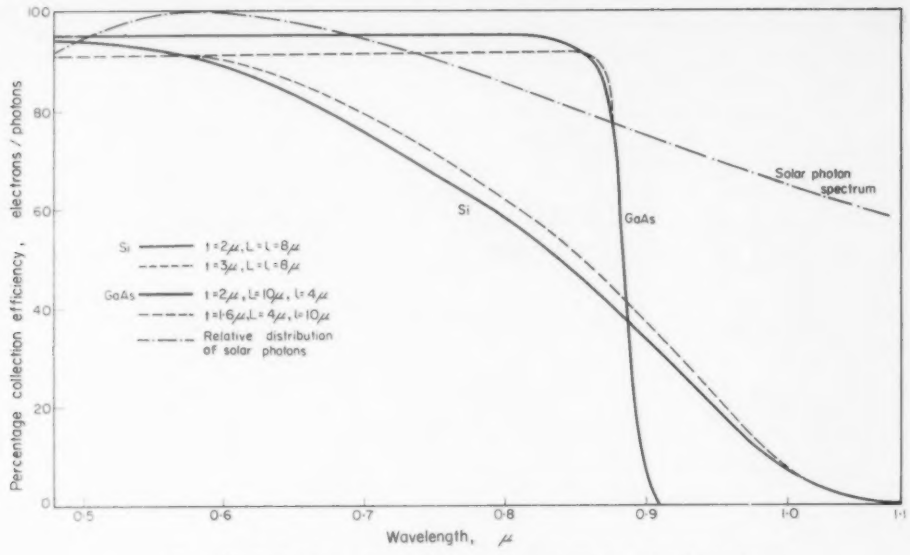


FIG. 4. Collection efficiencies for Si and GaAs solar batteries.

absorption has been included and no allowance has yet been made for surface reflection, i.e. the tabulated results are in terms of photons entering the semiconductor.

(The diffusion lengths taken are considered to be about the best obtainable for the doping levels used in Table 2.)

It will be seen from the table that the performance of the GaAs units is slightly superior.

4.2 *Intrinsic carrier concentration*

The significant advantage of the GaAs unit is that this somewhat higher collection efficiency is achieved with a room-temperature energy gap of $E = 1{\cdot}42$ eV, compared with $E = 1{\cdot}1$ eV for silicon, and hence with a far smaller j_0 value. The intrinsic carrier concentration varies exponentially with energy, i.e.

$$n_i^2 \alpha \exp(-E/kT)$$

where T is the absolute operating temperature, taken as 300°K.

This expression is thus $\sim 10^5$ times lower for GaAs than Si. The intrinsic carrier concentration

Table 2. Solar battery parameters

Parameter	Silicon	GaAs (normal)	GaAs (reversed)
Surface doping*	$3\times10^{18}p$	$5\times10^{17}p$	$5\times10^{17}n$
Bulk doping	$10^{18}n$	$5\times10^{17}n$	$5\times10^{17}p$
Surface thickness	2 or 3μ	2μ	1·6μ
D in bulk	7·5 cm^2/sec	7·5 cm^2/sec	60 cm^2/sec
D in surface	25 cm^2/sec	60 cm^2/sec	7·5 cm^2/sec
$\tau_h n_s = \tau_e p_s$	10^{11} cm^{-3} sec	10^{10} cm^{-3} sec	10^{10} cm^{-3} sec
Diffusion lengths:			
Surface	9μ	10μ	4μ
Bulk	9μ	4μ	10μ
j_0/n_i^2	$\sim 1{\cdot}1\times10^{-14}$	$5{\cdot}5\times10^{-14}$	12×10^{-14}
j_0	$4{\cdot}5\times10^5$	0·22	0·48

* Heavier surface doping has been assumed for the Si unit to obtain low transverse resistance.

also depends on the effective masses of the carriers, the complete expression being (at 300°K)

$$n_i^2 = (2\pi m_{\text{eff}} kT/h^2)^3 \exp(-E/kT)$$
$$= 6{\cdot}3\times10^{38}\,(m_{\text{eff}}/m_0)^3 \exp(-E/kT)\ \text{cm}^{-6} \quad (33)$$

where m_{eff} is an effective density-of-states mass.

For silicon, allowance for the six minima in the conduction band gives[7]

$$(m_{\text{eff}}/m_0)^3 = 0{\cdot}52$$

For GaAs, we have simply

$$(m_{\text{eff}}/m_0)^3 = (m_e^* m_h^*/m_0^2)^{3/2} \quad (34)$$

where $m_e^*/m_0 = 0{\cdot}072$ and $m_h^*/m_0 = 0{\cdot}68$ (EHRENREICH)[8] giving $(m_{\text{eff}}/m_0)^3 = 0{\cdot}011$.

Thus this density-of-states term is ~ 50 times smaller in GaAs than in Si, mainly as a consequence of the low effective-electron mass.

Hence at 300°K, we have for Si, $n_i^2 = 4\times10^{19}$ cm^{-6}, and, for GaAs, $n_i^2 = 4\times10^{12}$ cm^{-6} with an improvement in favour of GaAs of $\sim 10^7$:1.

The value of j_0 for Si is given in equation (32). For GaAs, with the present state of technology, lifetimes are worse than in silicon and they may well continue to be inferior because of the importance of direct recombination in GaAs.[9] We will, therefore, assume for GaAs $\tau_h n_d = \tau_e p_a = 10^{10}$ cm^{-3} sec, i.e. a factor 10 times lower than for Si, as calculated by MAYBURG[9].

The parameters taken for solar batteries of the two materials and the values of j_0 computed from equation (31) are given in Table 2.

4.3 *Expected efficiencies for* Si *and* GaAs *solar cells*

For solar radiation outside the earth's atmosphere (i.e. as for satellites) where the intensity is 0·13 W/cm^2, the photon flux is $Q_0 = 5{\cdot}8\times10^{17}$ photon/sec per cm^2.

Hence for silicon, from Tables 1 and 2, $J_0/j_0 = 4{\cdot}8\times10^{12}$ and, from equation (26), $r = 24$. For GaAs, $J_0/j_0 = (11 \text{ or } 4{\cdot}7)\times10^{17}$ and $r = 38$ or 37.

Thus from equation (28) the efficiencies are Si = 16 per cent, GaAs = 28 per cent (normal) and 26 per cent (reversed). In practice these values will be reduced by two loss factors.

(1) Reflexion at the front surface. The method of forming the p-layer in silicon junctions gives very low reflectivity,[10] and about 95 per cent of the radiation enters the semiconductor.

(2) Imperfect electrical characteristics of the unit. According to equation (26) the "junction efficiency" defined as the ratio of the optimum output power to the product of the open-circuit voltage and short-circuit current should be

$$\eta_J = W_{\text{opt}}/J_0 V_{0c} = \frac{r-1}{r+\log(r+1)} \quad (35)$$

$= 0{\cdot}84$ for $r = 24$ as above.

By comparison, the figure obtained from the current/voltage curves of the latest "gridded" Hoffman solar cells (of 13 per cent efficiency) is 0·72. Hence the junction efficiency, including contact losses, is about 86 per cent of theoretical. Combining these figures, the resultant efficiency expected for the best silicon solar cells is

$$\eta = 16\times0{\cdot}95\times0{\cdot}86 = 13 \text{ per cent}$$

This value is in excellent agreement with the performance obtained by the latest Hoffman cells. The limit of development of the silicon cell would seem to be little better than the above figure. It is unlikely that the absorption can be reduced significantly below 5 per cent already assumed, while a junction plus contact efficiency exceeding 90 per cent of theoretical seems unlikely. Thus the limit of the silicon solar battery is probably an efficiency of about 15 per cent, with production units inevitably somewhat worse than this.

The loss factors may be estimated also for GaAs units:

(1) It may be difficult to obtain quite such low reflectivities as for silicon, so we will assume the more pessimistic surface transmission factor of 90 per cent.

(2) The technology of GaAs is not yet as highly developed as for Si, so that at the present time p–n junctions in this material are far from ideal. However, there seems no reason why intensive development work should not remedy this situation, so that in the not too distant future junction efficiencies about 80 per cent of theoretical may be achieved for "normal" GaAs junctions.

With these parameters a realistic figure for the

expected efficiency of a normal GaAs solar battery is

$$\eta = 28 \times 0{\cdot}9 \times 0{\cdot}8 = 20 \text{ per cent}$$

For the "reversed" unit with a surface n-layer it should be possible to reach very high junction efficiencies, as the electron mobility is 2500 cm^2/V-sec even in material with 10^{18} cm^{-3} free carriers.[8] With good development work it would seem reasonable to hope for a junction efficiency of 90 per cent of theoretical, giving

$$\eta = 26 \times 0{\cdot}9 \times 0{\cdot}9 = 21 \text{ per cent}$$

The best GaAs units reported to date[11, 12] have efficiencies of 7 per cent. It has been suggested by LOFERSKI *et al.*[13] that recombination and regeneration of carriers in the depletion region is a major source of inefficiency. However, the results reported by both the above groups show that the open-circuit voltages are quite high ($> 0{\cdot}8$ V) which is nearly 90 per cent of that expected from the present theory, showing that the junctions are in fact rather good. The same data show that the major failing is in the short-circuit current, which is only half the expected value. This low collection efficiency stems from low diffusion lengths, which, in a 5 per cent efficient cell[13], were only $\sim 0{\cdot}2\,\mu$ and in the 7 per cent cell[11] had been increased to $\sim 0{\cdot}6$.

5. CONCLUSIONS

Detailed analysis of the photoelectric processes in p–n junctions has been used to estimate the probable efficiencies of silicon and gallium arsenide solar batteries.

The theory has shown that for any given absorption constant there is an optimum thickness of surface layer. In the past,[3] it has been considered that this layer would be indefinitely thin if it were not for the effects of lateral resistance in the layer. The optimum thickness is roughly the reciprocal of the absorption constant, averaged over the effective waveband involved. Surface recombination is not a very important factor, particularly when the condition of optimum thickness is realized, or approached.

The numerical results obtained show that silicon units cannot be expected to be much better than 13 per cent efficiency at the present time, even using the "gridded" structures to reduce the transverse resistance, and that the limiting efficiency achievable in the future will be about 15 per cent. A similar conclusion has been reached by WOLF[3].

By contrast, GaAs units, with moderate development, should reach 20–21 per cent efficiency, the higher figure being obtained for a reversed configuration, i.e. the surface layer n-type and the bulk-material p-type. The ultimate development limit should approach 25 per cent.

In order to achieve these results it will be necessary to advance the state of GaAs technology to approximately the level of present day work on silicon in respect of:

(1) Doping techniques
(2) Fabrication of sharp junctions
(3) Low resistance contacts
(4) Reflectivity and surface recombination

In addition it is very important to obtain carrier lifetimes as long as possible, but it has *not* been assumed in the theory that these will ever approach the values obtained for Si. On the contrary they have been taken as only 10 per cent of the values for Si of the same degree of doping. Also the reflectivity of GaAs has been assumed to be somewhat higher than Si.

The fundamental advantages of GaAs over Si are:

(1) GaAs has a very sharp absorption edge, as a consequence of direct transitions from valence to conduction band.
(2a) Activation energy of GaAs is considerably higher.
(2b) Lower effective masses and only one conduction band minimum, giving low-density states.
(3) Very high mobility giving low transverse resistance in the case of the "reversed" design.

The result of items (2a) and (2b) is that the minority-carrier concentration for a given doping level is $\sim 10^7$ times lower than in Si.

The only fundamental disadvantage of GaAs is that the lifetimes will probably be shorter than in Si because of the influence of direct recombination. Values only 1/10 of those for Si have been assumed in the analysis, and the effect of this is only to reduce the factor in (2) above by 5 or 10 times,

Vol. 2
1961

still leaving a net advantage to GaAs of $\sim 10^6$:1.

In general terms, a good GaAs solar battery should give a slightly higher short-circuit current than a Si unit, and about 50 per cent greater voltage. The state of GaAs technology at the present day is, of course, well below that of silicon, but the above figures show such superior performance for GaAs solar batteries that the necessary development work is clearly called for. As, in addition, GaAs is proving a very promising material for transistors and various types of diodes, there seems little doubt that devoting intensive effort to the technology of this material is justified.

REFERENCES

1. D. M. Chapin, C. S. Fuller and G. L. Pearson, *J. Appl. Phys.* **25,** 676 (1954).
2. T. S. Moss and T. D. F. Hawkins, *Infrared Phys.* In press (1961).
3. M. Wolf, *Proc. I.R.E.* **48,** 1246 (1960).
4. T. S. Moss, *Optical Properties of Semiconductors.* Butterworths, London; Academic Press, New York (1959–61).
5. W. C. Dash and R. Newman, *Phys. Rev.* **99,** 1151 (1955).
6. R. Braunstein, A. R. Moore and F. Herman, *Phys. Rev.* **109,** 695 (1958).
7. H. Brooks, *Advanc. Electron. Electron Phys.* **7,** 85 (1955).
8. H. Ehrenreich, *Phys. Rev.* **120,** 1951 (1960).
9. S. Mayburg, *Solid State Devices Conference, Carnegie Tech., Pittsburgh, June,* 1960.
10. J. F. Hall, *J. Opt. Soc. Amer.* **50,** 717 (1960); M. Wolfe and M. B. Prince, *Brussels Solid State Congress* p. 1180. Academic Press, New York (1960).
11. J. J. Loferski, P. Rappaport and J. J. Wysocki, *Proceedings of 13th Annual Power Conference.* U.S. Army Signal Research and Development Laboratory (1959).
12. E. G. Bylander, A. J. Hodges and J. A. Roberts, *J. Opt. Soc. Amer.* **50,** 983 (1960).
13. J. J. Wysocki, J. J. Loferski and P. Rappaport, *Proceedings of 14th Annual Power Conference.* U.S. Army Signal Research and Development Laboratory (1960).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 232–246. Printed in Great Britain

INJECTION ELECTROLUMINESCENCE

ALBRECHT G. FISCHER

RCA Laboratories, Princeton, New Jersey

(*Received* 7 *October* 1960; *in revised form* 8 *December* 1960)

Abstract—After a brief review of previous literature on the subject, it is shown that the theoretical limit for brightness and efficiency of injection EL can be expected to be higher than that of impact ionization EL. A survey of promising host crystal materials is given, and the unusual difficulties encountered in making junctions in wide-band-gap materials are pointed out. *p–i–n* Junctions offer a more favorable approach to efficient injection EL than *p–n* junctions. The properties of varied band gaps at the contacts will be described. A discussion of sources of loss will be given.

Résumé—Après un bref aperçu résumant les articles publiés ayant trait à ce sujet, il est démontré que la limite théorique pour l'éclat et le rendement d'injection électroluminescence (EL) peut dépasser celle de l'ionisation de choc EL. Une étude des matériaux de cristal hôte est décrite et les difficultés exceptionnelles que l'on rencontre dans la fabrication des jonctions dans les matériaux à espace de bande large sont particulièrement indiquées. Les jonctions *p–i–n* offrent une approche plus favorable à une injection EL que les jonctions *p–n*. Les propriétés de divers espaces de bande aux contacts seront décrites. Une discussion des sources de perte sera donnée.

Zusammenfassung—Nach einem kurzen Überblick über die existierende Literatur wird nachgewiesen, dass die theoretische Grenze für Leuchtkraft und Leistung bei Injektionsleuchten höher sein muss als bei durch Stossionisation erzeugtem Leuchten. Es folgt eine Übersicht über aussichtsreiche Wirtskristall-Materialien. Die aussergewöhnlichen Schwierigkeiten bei der Herstellung von Übergängen in Materialien mit breiter verbotener Zone werden erörtert. *p–i–n*-Übergänge sind zur Erzeugung von gutem Injektionsleuchten geeigneter als *p–n*-Übergänge. Die Eigenschaften von verbotenen Zonen verschiedener Breite an den Kontakten werden beschrieben und die Verlustquellen werden diskutiert.

INTRODUCTION

THE efficient generation of visible light by means of carrier injection and recombination in junctions of wide-band-gap semiconductors is an extremely fascinating idea.

This paper is the first attempt to give a review and a preview of this still embryonic field, with the objective to stimulate interest and thought.

The Destriau type of electroluminescence (EL) has been extensively investigated in the past decade but still resists complete understanding. Its practical applications are limited to low-brightness panels of high impedance for a.c. operation.

Contrary to this, the Lossev type of EL, which is our subject here, is still in an undeveloped state, but is simple to understand in its principles. It is nothing but a straightforward fusion of semiconductor and luminescence physics. Notwithstanding the experimental facts to date and despite the lack of a complete quantitative theory, there are good indications that injection EL has superior capabilities as an efficient, bright light source for d.c. and a.c. operation, and that it will be suitable for many important applications. The experimental difficulties which have yet to be overcome will be pointed out. Despite the fact that efficient injection EL will be easier to accomplish for infrared emission, our discussion will be mainly concerned with the visible spectral range, since a breakthrough here would be most significant.

Owing to the undeveloped state of this field it is often unavoidable that speculations, preliminary results and personal opinions take the place of theory and exact data. It is hoped that this is a

stimulus to the reader to find more accurate answers wherever possible.

1. COLLISION EXCITATION vs. INJECTION EL

A. *Impact ionization EL*

The electroluminescence of ZnS-type phosphor powders which are embedded into an insulating medium and exposed to an alternating electric field (DESTRIAU, 1937) is widely known.[1] It is generally ascribed to recombination of excess carrier pairs which have been generated by impact ionization in high-field regions in insulating barriers.* Such barriers are interspersed between conducting paths within the crystallites. The same type of EL also occurs in the high-field region of blocking contacts on conducting luminescent single crystals or sintered layers (Si, Ge, SiC, GaP, GaAs, ZnO, Cu_2O, AlN, CdS, ZnS). The collision–excitation mechanism requires that electrons of thermal energy gain further energy from the field up to an energy of several electron volts. The strong interaction of "warm" electrons with acoustical and optical phonons of the lattice and with crystal imperfections and impurities renders this process a very inefficient one. According to recent calculations by MOREHEAD[2], only one of every ten electrons that traverse a luminescent barrier manages to break through the "phonon friction barrier" and to acquire impact ionization energy. Other ohmic losses of intrinsic nature are caused by the charge transport that occurs within the conducting spikes and that is necessary to build up the high-field regions. Further ohmic losses are due to the back-flowing of impact-generated excess electrons to their original centers. Some of the ohmic loss is indispensable since the high-field effects in barriers have negative current–voltage characteristics and need current-limiting ballast resistors to prevent breakdown.

Most of the published computations of the maximum efficiency obtainable with this type of EL are generally too complicated to be convincing. A recent estimate[2] is 37 lm/W. However, since impact EL is a special form of low-voltage cathodoluminescence, a simple comparison between cathodoluminescent and electroluminescent mechanisms reassures the above-mentioned figure as an upper limit.

The brightness of EL powder cells is also limited. Since there is no direct supply of charge from the electrodes, the maximum obtainable brightness of the Destriau EL is limited by the small number of bound electrons available for ionization and subsequent acceleration. Besides, only a very small fraction of the total crystal volume becomes excited. Expensive high-frequency, high-voltage operation is required to obtain higher brightnesses. One can ask whether the properties of impact EL light sources can be improved by using more favorable physical arrangements. If one uses directly contacted binderless layers of EL materials, either as coherent sintered layers or thin single crystals, the operation with d.c. becomes possible. Through this the brightness may be increased. The efficiency is decreased, however, because the impact-generated excess electrons can never return to their native centers in this case, and the recombination cross-section of the ionized centers in the barrier for other electrons entering the crystal over the cathodic barriers is considerably lowered by the constant high field.*

A good arrangement for d.c. impact ionization and direct electroding would be a *p–i–n* junction biased in reverse, with an avalanche-forming high-field region both at cathode and anode. For a.c. operation, a *p–n–p* structure would be most favorable. Such structures are being investigated as a side-line in the course of this program. Because in this case junctions which are biased in reverse directions are being used, such structures

* Other possible carrier-generation mechanisms, which, to the author's mind, have not been sufficiently investigated, are: thermally aided Zener emission at imperfection-impurity clusters; stepwise, field-aided generation at impurity level "ladders" across the forbidden gap; field-aided thermal generation in low-band-gap inclusions with subsequent injection of both electrons and holes into the adjacent luminescent material and recombination there at the sites of trapped holes at field reversal. Carrier injection can also take place at the contact with the embedding medium which is an electrolyte.

Two-carrier, tunnel injection from surface states, with trapping of the injected carriers for half a cycle, also seems conceivable.

* *Note added in proof:* The last argument does not hold for luminescent centers of the manganese or rare-earth type, where impact-excited transitions occur without ionization of the center.

might be termed "reverse injection" EL light sources.*

B. *Forward-injection EL*

Luminescence in semiconductor-type phosphors like ZnS is caused by recombination of excess carrier pairs via luminescent centers. The excess carrier pairs can be generated by light absorption (photoluminescence) or by electron impact (cathodoluminescence, X-ray-luminescence and impact-electroluminescence). In all of these cases, indirect methods are necessary to produce the excess carrier pairs, (ultraviolet-photons, X-ray photons, cathode rays, hot electrons). Obviously, the electrical injection of excess carriers as conduction currents via contacts into the luminescent crystal would be much more direct. Such a light source would simply consist of a wafer of semiconducting luminescent material having an internal p–n junction and two wires attached to it (Fig. 1).† When biased in forward direction, electron and hole current overlap each other and recombination accompanied by light emission takes place.

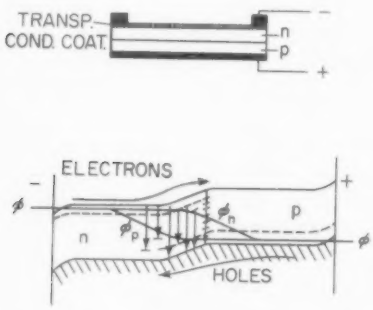


FIG. 1. Physical and band model of p–n-injection EL arrangement. Seven different possibilities for electron–hole recombination are indicated.

In the first, naïve approach which we pursue in this section, the prospects of such a light source are rather fantastic.

Assuming one has a p–n junction of 1 cm² area in a crystal capable of emitting green light, a green band like the luminescence of ZnS:Cu has an energy-equivalent of 450 lm/W and centers at $h\nu = 2{\cdot}5$ eV.

If a forward current of only 1 A/cm² flows through this junction (Si-diodes can easily handle 10^3 A/cm²) and if, as is the case in good ZnS phosphors, 60 per cent of the recombinations are radiative, then a light energy of $2{\cdot}5 \times 1 \times 0{\cdot}6 = 1{\cdot}5$ W/cm² = 675 lm/cm² is emitted from this junction, corresponding to a brightness of 6×10^5 ft-L (tungsten filament, 3200°K, 5×10^6 ft-L). It is known from pulsed, high-intensity cathodoluminescence that phosphors like ZnS can very well stand much higher excitation levels without decrease in efficiency if heating is avoided.[4]

Not only the brightness, but also the efficiency of such light sources may be very high. Good ZnS-type phosphors have energy conversion factors of 20 per cent for cathode-ray excitation and 40 per cent for ultraviolet excitation. There is no reason to believe that such high figures should be unattainable with direct electrical excitation. Additionally, because of thermal diffusion currents, the voltage that is necessary to drive a charge carrier across the junction is considerably less than the 2·5 eV carried away by the photon that is created at the sign reversal of the charge carrier. LEHOVEC *et al.*[5] observed yellow light emission ($h\nu = 2$ eV) from SiC junctions with only 1·8 V applied. Thus the light source absorbs heat from its surroundings and converts phonons into photons. Theoretically, the ratio of light output to electrical input energy can become greater than one. This holds at least for low currents (and brightnesses) and elevated temperatures where the contribution of thermal-diffusion currents is not negligible against field-driven currents. If the temperature is too high, radiationless transitions take over and the efficiency drops. A thermo–opto-electrical cooling device based on this principle is infeasible because at reasonable electrical input power joule heating offsets the cooling effect.[6]

2. BRIEF HISTORY OF INJECTION EL

Light emission in rectifying SiC crystals with a

* The strong connection between this type of EL and cathodoluminescence suggested the investigation of a flat cathodoluminescent lamp. Since in this case the electrons are accelerated loss free in a vacuum, higher brightness and efficiency than in EL might be possible. Previous work of the author confirmed that such a lamp using a large-area, self-sustained MgO cold cathode is feasible.

† However, light sources employing microcrystalline powder particles with intergranular contacts are equally possible.[3]

Vol. 2
1961

forward bias applied was found by LOSSEV in Russia in 1923.[7] In the early thirties, LOSSEV attributed the effect to visible bremsstrahlung. He recognized that this crystals consisted of *n*- and *p*-type regions, separated by a highly resistive interlayer, but he lacked the concept of carrier injection.

In 1948, TETZNER[8] concluded that the light emission from SiC crystals must be luminescence since it had the same spectral distribution as photoluminescence and cathodoluminescence of the same crystals. After the phenomenal surge of semiconductor physics between 1948 and 1951, LEHOVEC *et al.*[5] explained the phenomenon in terms of *p–n* injection and radiative recombination in forward-biased *p–n* junctions.

PATRICK[9], RÜCKER[10] and the present author[11] rediscovered that the junctions were not *p–n*, but *n**–*n*–*p** junctions with a highly resistive *n*-layer between highly conducting *n**- and *p**-layers. This author also found green and red emitting junctions as opposed to the broad yellow band so far reported.[11] Using a boron–aluminum mixture, the author prepared an artificial luminescent junction on SiC,[11] a technique that was later perfected by HALL[12].†

Other materials than SiC have been reported to display forward injection EL, however, in the infrared region. In the early fifties, forward *p–n* injection was found in germanium and silicon by HAYNES and BRIGGS[13], and NEWMAN[14]. Both band-to-band and impurity transitions were observed. In InSb, the effect was reported by MOSS *et al.*[15] In GaAs, GaSb, InP and InAs, it was observed by BRAUNSTEIN[16]. VAN DOREN and DE NOBEL found it in CdTe.[17] Recently, the phenomenon has been observed in the visible and infrared spectral range in GaP by LOEBNER and POOR[18] and by FROSCH and GERSHENZON[19].‡ The possibility of the conversion of heat into light in a wide-band-gap *p–n* junction first mentioned by LEHOVEC[5], was further substantiated under idealized thermodynamic asumptions by WEINSTEIN[20]. WILLIAMS[21] claimed that irreversible thermodynamics forbids efficiencies greater than 100 per cent.

To this author, this whole question seems to be purely academic, because there are other unavoidable sources of losses which keep the maximum efficiency in the order of 10–20 per cent [see 4(F)].

The possibility of using *p–n* injection EL in connection with photoconductors for relays, switches, electrical and optical amplifiers has been pointed out by LEHOVEC[22], LOEBNER[23] and DIEMER[24]. The advantages of *p–i–n* injection over *p–n* injection was outlined by the author in his thesis,[11] in a German Patent,[25] and at the Garmisch conference[3]. SMITH[75] observed two-carrier injection EL in CdS.

3. A SURVEY OF HOST CRYSTAL MATERIALS

Though the effect was discovered in SiC, this material did not appear to this author as a very suitable starting material. The properties of SiC (melting point higher than 2900°C, decomposition above 2200°C, extreme hardness, impurity diffusion not below 1800°C, about thirty different modifications, poor luminescent properties) led to the question of whether there are other, better-manageable materials available to study the effect.

Suitable materials have to meet the following criteria:

I. In order to allow optical transitions in the visible spectral range, the material must have a band gap of larger than 2·5 eV. It must have yellow or white body color.

II. The material must show efficient luminescence of the free-carrier type with photo- or cathodo-excitation. This means that either radiative band-to-band transitions must be possible, or, in the case of impurity transitions, the activator level must be more than 0·4 eV apart from its native band to prevent thermal ionization at room temperature.

III. In order to be capable of amphoteric semiconduction with high and approximately equal carrier mobilities, the material must have wide allowed bands, which is only possible with strongly covalent bonding. The material must have diamond structure or a structure derived thereof (zinc blende, wurzite). The amount of covalent vs. ionic bonding can be estimated from the electronegativity values of the constituents.[65] From this it follows that III–V compounds have preference.

† *Note added in proof:* Recently, injection EL from grown *p–i–n* junctions in SiC was reported.[28]

‡ *Note added in proof:* GRIMMEIS *et al.* observed injection EL in AlP,[67] GRIMMEIS and KOELMANS in GaP.[72]

Binary compounds should have preference over ternary or even more complicated compounds to avoid unnecessary trouble.

IV. The properties of the material must be such that thin crystal wafers can be prepared from it with reasonable technical effort. Preferably, pulling of crystals from the melt and subsequent slicing as it is practiced with semiconductor materials should be possible. Vapor-phase reaction and chemical-transport methods are generally not so desirable because of low yield and difficult control. However, if they can give near-single crystalline layers at low-substrate temperatures they may become important.

These criteria allow us to list the compounds in Table 1.

Table 1

IV–IV	III–V	II–VI	I–VII
Diamond SiC	BN AlN GaN BP AlP GaP BAs	ZnO CdS ZnS ZnSe (ZnTe)	CuI CuF (Alkaline-earth chalcogenides have not yet been studied because of chemical instability)

To evaluate these materials, an extensive study has been and is being carried out by the author. The results can be briefly summarized as follows.

Diamond

The EL of diamonds was studied by the author and others in 1957.[26] Bombardment of diamond surfaces with boron ions produced hole-injecting contacts on insulating crystals. Diffused contacts should be possible with the autoclaves now available which make it possible to heat diamonds without conversion into graphite. With such means, it would also be possible to diffuse activator centers into non-luminescent diamonds. As a purely homopolar crystal, diamond offers the unique advantage of high and approximately equal carrier mobilities.

Of course, cost considerations and the difficulty of artificial synthesis of diamonds preclude practical applications, at least at present.

SiC

Some results with this material were mentioned before. The crystals were prepared using the sublimation method of LELY[27]. Though alloyed contacts could be prepared, the technology of this material seemed too difficult in comparison with other materials mentioned below. Recently, EL from grown *p–i–n* junctions was reported.[28]

BN

Boron nitride requires about 40,000 atm for the conversion into the cubic phase.[29] The possibility for this conversion was predicted by the author,[3] and a suitable autoclave was suggested in a patent.[30] The prospects to produce *p–n* junctions in such crystals in a controlled manner are very remote at present.

BP

This material was prepared by reaction of PH_3 and BBr_3 at 1200°C in a quartz tube. It came in the form of black, hard crusts of *n*-type conductivity and very good chemical stability.[11] Judged from electronegativity values, the compound is nearly as covalent as diamond, more covalent than SiC. The bandgap is in the neighborhood of 5 eV. The melting point should be about 2500°C, requiring a very high phosphorus pressure to prevent decomposition. Therefore, growing from the melt seems to be impractical. Recently, *n*- and *p*-type material has been prepared by vapor-phase reaction and by crystallization from an undisclosed flux.[31]

BAs

This material has been prepared,[32] but not much is known about it. It can be expected that its bandgap falls in the visible range.[66] It should be

Vol. 2
1961

practically completely covalent, which would give high and equal carrier mobilities. Its usefulness for purposes of electroluminescence was proposed in Ref. 33.

AlN

Contrary to early reports in the literature which are based on extrapolation, the author found that this substance has its melting point over 2900°C. *n*-Type conductivity and impact-ionization EL were observed in vapor-phase grown crystals.[11,64] For injection EL, the material seems unsuitable because of its wide bandgap (5 eV) and its high degree of ionicity.

AlP

AlP is a highly covalent material with a bandgap of about 2·5 eV. Small crystals were prepared from Al and P using bromine as a carrier gas. The crystals are unstable against moisture which reduces their practical usefulness. Preliminary results indicate that they can be obtained *n*- and *p*-type, insulating and luminescent. It can be predicted that melt-growing will be difficult. Transport reactions and growing from a flux is more promising. Recently, injection EL, photoconductivity, photovoltaic effect and *n*- and *p*-type conduction of AlP were found by GRIMMEIS *et al.*[67]

GaN

Parallel to and confirming recent investigations by others[35, 36, 34b], GaN was prepared by the author in powder form by reaction of gallium with ammonia and subsequent removal of excess Ga with halogens, by reaction of a Li–Ga-alloy with ammonia whereby the Li distills off as $LiNH_2$, by decomposition of $(NH_4)_3GaF_6$ in ammonia, by blowing a Ga-mist into hot ammonia and by heating in ammonia of an ultrasonically prepared emulsion of liquid Ga in petroleum jelly. The material is luminescent[36] and can be made *n*- and *p*-type. It shows band-to-band recombination radiation in the ultraviolet,[36] even at room temperature. Thermal decomposition sets in at 950 °C. This temperature is so low that the diffusion of impurities into the lattice is still impossible. The estimated melting point is 2000 °C. Attempts to heat the material undecomposed under high nitrogen pressure have failed. Transport reactions and growing from flux are the only ways to get crystals.

GaP

Gallium phosphide is a very suitable material. Impurity-center transitions cause deep-red and near-infrared emission. Band-to-band recombination occurs in the green, even at room temperature. GaP can be made *n*- or *p*-type or insulating. The ratio of electron-to-hole mobility is favorably low. Recently, single crystals have been grown from the melt with RF heating under 25 atm of phosphorus pressure.[37] Diffused Zn contacts on *n*-type crystals produced forward and reverse injection EL.[37] The author found an easy way to prepare GaP powder in large quantities by conversion of GaN in phosphorus vapor.[11, 38]

II–VI compounds

The zinc and cadmium chalcogenides seem to be very promising compounds for our purpose, and most of the following discussions will concentrate on them. They are known as the most efficient phosphors and photoconductors, and can be prepared as single crystals and thin layers by several different methods (though not very easily). The general rule holds that in II–VI compounds the tendency to become *n*-type increases with the size of the cation, whereas the tendency to become *p*-type increases with the size of the anion (see Table 2).

*Table 2**

	Zn	Cd	Hg	→*n*
O	ZnO	(CdO)	(HgO)	
S	ZnS	CdS	(HgS)	
Se	ZnSe	(CdSe)	(HgSe)	
Te	(ZnTe)	(CdTe)	(HgTe)	
↓ *p*				

* Compounds in brackets are infrared emitters.

ZnO

Zinc oxide has not yet been grown from the melt because it requires high oxygen pressure to prevent decomposition and because no existing crucible material withstands its chemical attack.

Vapor-phase growth is rather troublesome. Recently it was found that ZnO platelets can be grown out of a PbF_2 flux.[39] ZnF_2 is probably even better suited. Hydrothermal methods have rendered big crystals.[69]

A slight excess of Zn makes the crystals green luminescent and *n*-type conductive. *p*-Type conductivity by Li-additions has been reported,[40] but occurs only at very high temperatures. Reverse-injection EL was found by the author in 1953.[41] *p*-Type CuI and Cu_2O contacts on ZnO displayed no forward injection EL as yet.

ZnS, ZnSe, CdS

Previous work[42–44] has established that these compounds can be crystallized from the melt. The equilibrium pressures to maintain stoichiometry are unexpectedly low (1·5–7 atm). The melting points are 1830°C for ZnS,[42a] 1515°C for ZnSe,[44] and 1485°C for CdS.[42b] The author developed techniques for RF zone melting and Czochralsky pulling[45] under controlled, pressurized atmospheres. ZnS is less suitable because it has two modifications which lead to stacking faults, a band gap of 3·7 eV which is too wide, and because it is too difficult to make conducting. ZnSe (2·7 eV) is single-phase cubic (the tendency to become hexagonal below 350°C is insignificant). CdS (2·5 eV) is hexagonal, with an insignificant cubic range at low temperature. ZnSe–ZnS and CdS–ZnS solid solutions are single-phased over a wide range of compositions, which makes it possible to vary the band gap to a wide extent.

ZnSe and CdS form a continuous series of quasibinary solutions which are hexagonal for CdS contents from 100 to 30 per cent and cubic for ZnSe contents from 70 to 100 per cent.[70] CdS and ZnTe have a wide miscibility gap.[70] ZnSe can be made *n*- and *p*-type[44] though not very strongly. CdS and CdS–ZnS mixtures rich in CdS can be made strongly *n*-type[46] without losing too much of their luminescence. This is due to the occurrence of very shallow donor centers in the Cd-sublattice. CdS with very high copper doping is *p*-type, but black. μ_p is 20 cm^2/V-sec.[47] The diffusion length of holes has been measured to be 3 μ,[73] a figure which is probably too low

ZnTe

ZnTe (band gap 2·2 eV) is normally *p*-type because it has shallow acceptor centers, but deep donor centers. It shows visible and infrared luminescence. It is most suitable for the preparation of hole-injecting contacts (see Fig. 5). It forms a continuous series of mixed crystals with ZnSe.[48] ZnTe–ZnSe solid solutions rich in ZnTe are still strongly *p*-type, but have a different band gap. ZnTe crystals can be grown from the melt and by sublimation techniques. Recently, alloyed *p–n* junctions have been formed in this material.[68] The hole mobility of ZnTe is 200 cm^2/V-sec.[74]

I–VII Compounds: CuI

Copper iodide has a strongly covalent zinc blende lattice and a band gap of about 3 eV. Crystals were prepared by sublimation, but growing from the melt (m.p. = 605°C) is also feasible. Crystals with excess copper show orange luminescence. Excess iodine causes discoloration and *p*-type conductivity. *n*-Type conductivity has not yet been achieved. *p*-Type CuI brought in contact with *n*-type CdS produced a junction with a photovoltage of 0·5 V. The material seems to be of potential value for making hole-injecting contacts on other crystals. A eutectic with CuCl which is also electronically conducting and *p*-type, melts as low as 284°C.

A drawback of CuI is its electrical instability due to evaporation of excess iodine, and its appreciable ionic conductivity in high fields. The latter leads to rapid deterioration of junctions in CuI. Despite this tendency, CuI might be very suitable for studies at low temperature.

Incidentally, CuI was one of the first semiconductors studied. Important properties of semiconductors (e.g. temperature dependence of conductivity, dependence of conductivity on doping concentration, positive Hall effect, etc.) were discovered and correctly explained in CuI by a forgotten pioneer of semiconductor physics, BAEDECKER, in 1909.[49]

CuF

Copper fluoride has a zinc blende lattice and can be expected to be more stable than CuI. However, the compound has not been prepared in crystalline form because of chemical difficulties. It is less covalent than CuI.

4. DESIGN CONSIDERATIONS FOR INJECTION EL LIGHT SOURCES

A. *Vacancy compensation of doping impurities*

In wide-band-gap semiconductors, we encounter a phenomenon which is unknown in conventional semiconductors. A new dimension of difficulty is introduced by the effects of "vacancy compensation".[50] Example I: If one adds aluminum to ZnSe to form donors, the crystal readily forms a zinc vacancy (which is a double acceptor) for every two aluminum atoms incorporated, so that the crystal remains insulating. Example II: If one adds copper to ZnSe to be incorporated at Zn sites to form acceptors, the crystal readily generates selenium vacancies (double donors) or interstitial copper atoms (donors) in the proper amount to keep the crystal insulating.

The reason for this behavior is simple: the crystal strives to get into the state of lowest energy. If the energy that could be gained by the transition of a newly introduced donor-electron into a virtual acceptor center is larger than the energy that is required to put this center into existence, then a vacancy which constitutes an acceptor center is formed so that the transition may take place.

Conversely: if a new acceptor center is introduced into a crystal and the energy that could be gained by the transition of a (not yet existing) conduction electron into this acceptor center surpasses the energy that is necessary to create the conduction electron, then the crystal readily generates donor vacancies to provide conduction electrons. Since the energy to form vacancies decreases with increasing temperature, this effect takes place at the high temperatures during preparation of the material.

It is evident that the tendency for vacancy compensation increases with the width of the band gap, and is inversely proportional to the depth of the incorporated center, which means that shallow centers are harder to realize than deep ones. Also, the formation of vacancies is easier in ionic compounds (e.g. I–VII compounds) than in covalent compounds (e.g. IV–IV compounds).

To minimize this undesired effect, we have to select materials with the smallest band gap possible. For instance, if orange EL is desired, we should not take ZnS (3·7 eV) but CdS (2·4 eV) or ZnSe (2·7 eV). Silver (shallow acceptor) is more difficult to incorporate in an electrically active form than copper (deep acceptor); aluminum (shallow donor) is more difficult to incorporate than indium (deep donor).

According to the mass-action law, the formation of vacancies can be partially suppressed by application of a pressurized atmosphere of the constituent that is being expelled from the crystal.[50] *n*-Type ZnSe or CdS crystals must be prepared or re-annealed in Zn or Cd vapor; *p*-type crystals must be prepared or re-annealed in selenium or sulfur vapor of elevated pressure. Conducting crystals must not be heated too high in neutral atmospheres because of resistance increase due to vacancy formation. These requirements add considerable difficulties to the preparation of junctions in wide-band-gap semiconductors.

B. *Problems caused by deep-lying levels*

In conventional semiconductors where donor and acceptor levels are very shallow so that they are thermally ionized at room temperature, the problem very often consists in increasing the resistance. In wide-band-gap materials, e.g. ZnS or ZnSe, the reverse problem exists.

In ZnS and ZnSe, the shallowest donor centers available are about 0·25 eV below the conduction band, whereas the shallowest acceptor centers are about 0·4 eV above the valance band.[76] At room temperature, only a small fraction of these centers is ionized. In order to get conductivities in the order of 10^4 Ω-cm, which is the minimum for efficient *p*–*n* injection EL, the concentration has to be raised to 10^{18}–10^{20}/cm^3. Because of the aforementioned vacancy compensation which cannot be completely suppressed, even higher concentrations are required. Such high concentrations of impurities frequently lead to discoloration of the crystals and self-absorption of luminescence. The interaction between activator centers, and the absorption due to trapped carriers, introduce radiationless transitions. These shortcomings are more pronounced with *p*-type doping than with *n*-type doping, because acceptor centers are deeper than donor centers. Since the luminescent transition usually consists of a free electron falling into a non-ionized (hole-occupied) acceptor centre, nearly all phosphors become nonluminescent if they become *p*-type.

ZnO, and CdS or CdS/ZnS-mixtures rich in CdS are the only phosphors which can be made strongly *n*-type without sacrificing too much of the luminescence. This is due to very shallow donors (0·03 eV) which occur in these crystals. ZnSe can be made *n*-type by strong doping with group III or group VII elements without quenching the luminescence.[44] These materials are, therefore, suitable for *p–n* injection EL.

C. *Luminescent transitions in p–n junctions*

As shown in Fig. 1, several kinds of transitions are possible. Band-to-band transitions would be very desirable if they were efficient enough at room temperature. They become efficient only in very pure crystals and with extremely high excitation densities, corresponding to carrier densities higher than $10^{19}/cm^3$.[51] At very high densities, light absorption due to free carriers forms a limit. Band-to-band transitions may be more predominant in crystals where the band extrema occur at the same *k*-value, making allowed transitions possible. With the exception of InSb where band-to-band transitions prevail, in most known cases of efficient luminescence the transition consists of a free carrier falling into a non-ionized activator center. Therefore, we will concentrate on this latter case here, though we keep the other case in mind. Usually, the free carrier is an electron in the conduction band, and the activator is a deep, non-ionized acceptor level above the valence band. Therefore, luminescent materials should have deep acceptors. If the acceptor level is ionized, either by elevation of temperature or because it is too shallow, a "radiationless transition" occurs. This can be caused by one of the following two mechanisms:

In both cases the primary transition is a band-to-band transition. In the first mechanism, the energy that is set free by a band-to-band transition can be transferred to a free or trapped electron or hole in the close vicinity, whereby this carrier is lifted high into its band and falls back under phonon emission (Auger recombination).[52] In the second mechanism, the energy released by the recombination can travel within the crystal, either as an exciton or by a repeated band-to-band-recombination–reabsorption–generation process, until the energy is captured by a "killer center" which converts the energy into many small quanta.

In the same host crystal, normally only one kind of luminescent transition exists, either from the conduction band into an acceptor center (Schön-Klasens model), or from a donor center into the valence band (Lambe–Klick model). The latter kind is the rarer. In very few phosphors has the simultaneous existence of both kinds been demonstrated.[77] Therefore, in a homogeneous *p–n* junction, only one branch of the junction, usually the *n*-type side, is capable of EL. All the carriers injected into the other branch perish there without radiation.

D. *Problems connected with p–n junctions*

If one wants visible emission via luminescent centers, ZnSe, CdS and CdS/ZnS mixtures, and ZnO are available as *n*-type crystals. In these materials, the electron mobility is much larger than the hole mobility. Usually emission takes place in the *n*-type branch of the junction. One naturally tries to make a strongly asymmetric junction (see Fig. 2). This requires an efficient hole-emitting contact. An ideal hole-emitting junction should inject holes far into the *n*-type crystal but should prevent electrons from leaving the *n*-type crystal.

There are three basic requirements for a good hole emitter: (1) It should have a higher absolute carrier concentration than the material into which it injects. (2) The hole mobility should be higher than the electron mobility. (3) The hole emitter should have a higher band gap than the material into which it injects in order to make it electron-repellent[53] (see Fig. 2).*

All three of these requests cannot be simultaneously fulfilled with presently known materials. Contrary to what one might wish, most wide-band-gap materials can more easily be made *n*-type than *p*-type. When they can be made *p*-type, then they are highly resistive and nonluminescent.

Unfortunately, in most materials the hole

* "Graded-seal" *p–n* junctions, made, for instance, from Ge–Si or Te–Se alloys, have already been proposed by LEHOVEC[54] for the purpose of facilitating the escape of intrinsic recombination radiation. LEHOVEC did not mention the other advantage of graded-seal junctions, namely the increased injection efficiency.

mobility is only a small fraction of the electron mobility.

The band-gap situation is also unfavorable. With the exception of diamond, SiC, GaN, CuI and possibly AlP and BP, p-type semiconductors

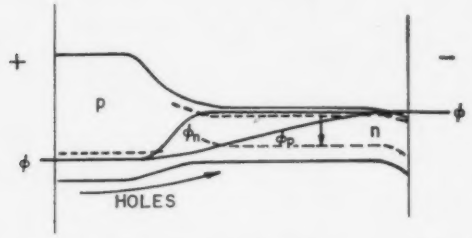


FIG. 2. Band scheme of an efficient assymetric p–n junction for forward-injection EL. The spread band gap of the anode repels electrons, preventing loss due to electron currents tailing into the anode and recombining there radiationless.

have narrower band gaps than GaP, ZnO, CdS and ZnSe. Thus, they are good "electron extractors", but inefficient hole injectors (see Fig. 3). Everything is just the reverse from what it should be. Inefficient injection at the p-type contact is one of the reasons why all p–n EL light sources so far investigated have very low efficiencies.

The conclusion must be drawn that, owing to a shortage of suitable materials, efficient p–n injection EL in the visible spectral range is not very easy to obtain. What is needed is a luminescent material for amphoteric conduction, or at least a

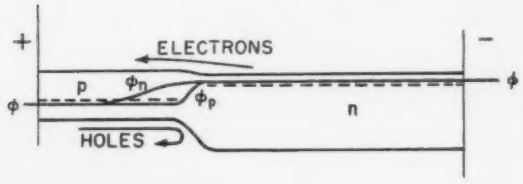


FIG. 3. Example of an inefficient hole emitter. If voltage is applied, more electrons will be extracted than holes emitted.

wide-band-gap, p-type conducting, luminescent material with manageable properties, and an n-type phosphor with high hole mobility. Both materials should be chemically similar so that solid solutions with variable band gaps could be

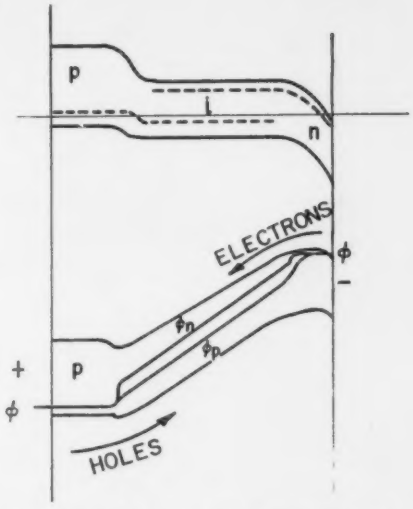


FIG. 4(a). Model of electroluminescent p–i–n junction in neutral state and with forward voltage applied. Note the spread band gap at the anode to repel electrons.

formed to make efficient injecting contacts. For infrared emission, the situation is much more favorable. Materials like CdTe, CdSe, ZnTe, GaAs, InP, GaP, HgS, PbS, Cu_2O, and combinations thereof, fulfill the above requests. It is

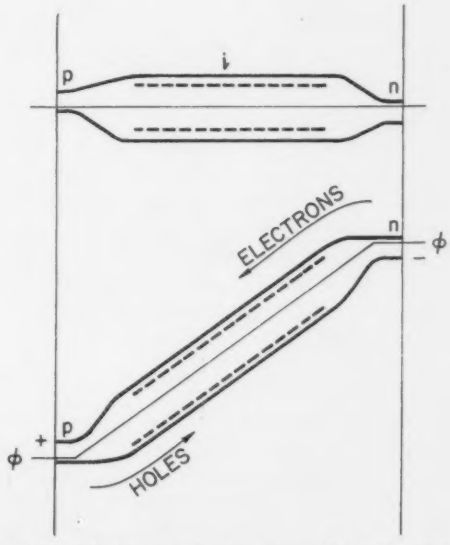


FIG. 4(b). Model of p–i–n junction with reduced band gap at the electrodes to facilitate injection.

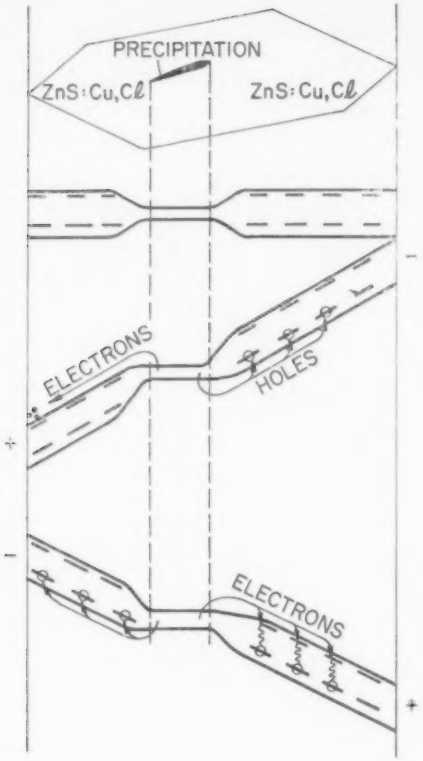


FIG. 4(c). Injection EL in microcrystallites embedded in an insulating medium.

probably much easier to solve our problem for infrared emission.

An additional severe shortcoming of forward-biased p–n junctions is the fact that recombination takes place within only an extremely thin layer.[71] Since we desire excitation of large volumes of crystals, we have to find ways to spread the thickness of the recombination zone, which leads us over to the next section.

E. *p–i–n Junctions*[11, 25, 3]

Instead of injecting minority currents into a semiconductor, there exists the other possibility of injecting diffusion or space-charge currents into an insulator or near-insulator (see Figs. 4 and 5).

This concept offers several decisive advantages: we can use a thin insulating wafer that is doped for optimum recombination efficiency without having to compromise with conduction problems. Into this wafer, electrons and holes can be injected from opposite surface contacts thus keeping semiconducting and luminescent parts strictly apart. Ohmic, electron-injecting contacts (cathodes) can be easily made,[78] even in transparent form[55–57]. The problems of the hole-injecting contact are considerably eased by the fact that, at least for low-injection current levels, the conductivity of the p-type surface layer need not be very high.

Such p–i–n junctions are also easier to fabricate than p–n junctions. In a p–n junction, the formation of the p–type electrode on the n–type crystal which usually requires heat treatment, impairs the n-type conductivity wholly or partly, either because of overlap of opposite doping regions and formation of neutral pairs, or by vacancy formation. Mostly one ends up with an involuntary p–i–n junction.[37, 28] Of course it is better to aim at an optimized p–i–n junction from the beginning. Naturally, equal and large carrier mobilities would be desirable also for the p–i–n case. Since strong space charges in the i-layer keep the concentrations of both carriers approximately equal, an electron mobility which is 10 times higher than the hole mobility would mean that 90 per cent of the electron current rushes into the anode unrecombined. Countermeasures to prevent this are: spread band-gap emitters[53] [see Fig. 4(a)], lowering of the Fermi level towards the valence band to improve hole injection, and a slightly blocking cathode to restrict electron injection. The thickness of the compensated or intrinsic i-layer plays a very important role. Of course one wants it as thick as possible in order to have more volume to be excited for light emission. On the other hand it is evident that the electrical resistance sets an upper limit. If the i-layer is very thin, in the order of the diffusion length of the slowest carrier, ($L_p \simeq 5$–$10\ \mu$ in CdS and ZnS), then space-charge restrictions due to injected free carriers or trapped carriers do not exist. The situation is described by existing theories on p–i–n junctions[58] which assume that the currents are diffusion currents. The forward current–voltage characteristic is given by an expression of the type

$$J \sim \exp\left(\frac{eV}{2kT}\right)$$

However, this theory is not immediately quantitatively applicable for our special case since it is tailored for the case of germanium where no deep nonionized traps are present, where there is an appreciable intrinsic conductivity in the *i*-layer, and where the mobility and diffusion length of injected carriers is very high compared to our materials. A rough computation, using proper adjustments, gave the result that, with a layer thickness of $10\,\mu$, a free-carrier-lifetime of 5×10^{-6} sec (taken from the decay-time of high-intensity cathodoluminescence) and an intrinsic carrier density of 10^6 cm^{-3}, the current rises from 1 A to excessive values between 1 and 2 V applied. Ohmic losses have been neglected in this theory.

If the *i*-layer is more than several diffusion lengths thick so that the current tails in front of each electrode do not reach the counterelectrodes, then the theory of two-carrier, space-charge-limited injection currents in insulators[59] applies. Here, the current is field-driven as opposed to the diffusion currents in the previous case. A theory describing the transition between the two cases is being developed.[60]

The simplified theory of LAMPERT[59] gives expressions for the current–voltage characteristic of the type

$$J = \text{const.}\,\frac{V^n}{L^m}$$

with L being the thickness of the *i*-layer, V the applied voltage, $n = 1/2 \ldots 3$, $m = 3/2 \ldots 5$. The values of n and m vary depending on the assumptions that are made on lifetime, mobility and residual conductivity of the *i*-layer.

It is important that this theory predicts that currents of 100 A/cm^2 can flow through insulating wafers of about 1 mil thickness with only 10–20 V applied.

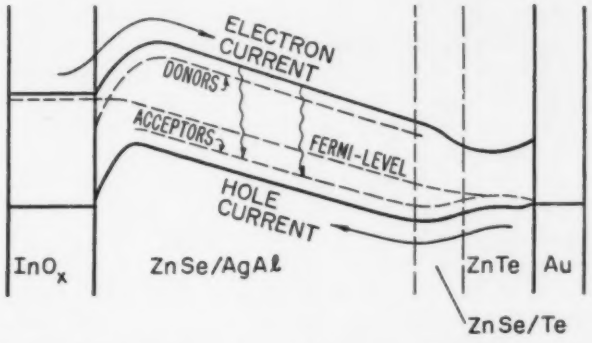


FIG. 5. Practical approach to injection EL. The figure shows a band model of a new light source.

The strong dependence of current on the thickness of the layer allows us to prepare structures for extremely high or extremely low current density to suit our purpose.

The theory quoted above has been derived for the case of diminishing recombination within the *i*-layer. For strong recombination, as it would occur in a material doped with 10^{18}/cm^3 luminescent donor–acceptor pairs, an amendment of the present theory is necessary which is being worked on. Also the effect of electrodes and the influence of the band gap has been neglected so far.

Further, the present theory holds only for the trapless case, for the case of very shallow traps, or for the case where the density of injected free carriers exceeds the density of trapped carriers.

If we can use materials with efficient band-to-band recombination (possibly GaN, ZnO, CdS, GaP, AlP), then we can verify the trapless case. For impurity-center luminescence, the trapped space charge in front of each electrode prevents further carriers from entering until the voltage has been increased far enough so that both space-charge clouds overlap and cancel each other. Then the current suddenly rises by several orders of magnitude up to the value prescribed by the

theory for the trapless case. The jump is often accompanied by negative resistance.[25,79]

This threshold-phenomenon may be desirable for certain applications; for others it is very disturbing. It can be eliminated by several measures:

(a) The *i*-layer can be made with a slight residual conductivity. Then there is, at low voltages, an ohmic region with light emission from a *p–n* junction. At higher current levels the space-charge-limited case is approached.

(b) By illuminating the crystal, free carriers are generated which compensate the space charges and thus initiate the current flow. To accomplish this in a practical device, the crystal could, for instance, be embodied in a Destriau-type EL matrix which gives a light flash upon application of the d.c. voltage. Incidentally, the self-induced photoconductivity of an electroluminescent *p–i–n* array due to reabsorbed emission is probably one of the reasons for the negative starting characteristics.

(c) Less practically, the starting could be accomplished by initial heating.

(d) A voltage pulse could start the device just like a fluorescent lamp.

(e) If one makes the *i*-layer thin enough so that one operates at the borderline between the space-charge-limited and the diffusion case, then the threshold disappears.

Another difficulty which is specific for injection into wide-band-gap insulators has never been explicitly mentioned but is always hidden in the postulate that the electrodes be "ohmic". The step in potential between the band edges of the highly conducting surface layer and the wide-gap insulator is nearly as large as half the band gap of the insulator, i.e. more than 1 eV. Very few carriers can overcome this step thermally. The field strength required to decrease the height of this barrier and to allow tunneling, would be so high that impact ionization phenomena would occur. Therefore it is essential that there is not an abrupt but rather a gradual transition in the conductivity from the surface electrode to the insulator.

If such soft transitions cannot be made, it is necessary to use reduced-band-gap contacts as shown in Fig. 4(b). These contacts have, of course, the disadvantage that unrecombined currents leak into the opposite electrode, especially so with strongly differing mobilities. Unless the *i*-layer can be made thick so that recombination in the electrodes becomes negligible as compared to volume recombination, this structure is, therefore, unfavorable as far as luminescent efficiency is concerned. If two reduced-gap contacts are used, the emitted radiation cannot even leave the crystal. Despite these shortcomings, a structure having a reduced-gap anode has been suggested in Fig. 5.

In this connection the following model which would make an injection EL powder cell for a.c. operation must be rementioned (see Section 1A, footnote, and Ref. 54).

Inclusions of spikes of low-band-gap, intrinsically conducting material in wide-band-gap, insulating EL powder particles, which are embedded in a transparent plastic, could be injecting contacts for electrons and holes into the insulator. A gradual transition from the low band gap to the high band gap would facilitate ohmic injection and make avalanche injection unnecessary. Electrons and holes are injected in opposite directions. The holes are caught by deep acceptors (activator centers) and recombine with free electrons after field reversal. [See Fig. 4(c).]

The existence of dark, elongated precipitations in copper-doped electroluminescent ZnS phosphor particles can be visually observed with a microscope[61] if the powder is embedded into a medium of high refractive index.[62] EL is emitted at the tips of these spikes. Unless the proposed scheme is already in existence in conventional EL cells, one should actively try to verify it. The advantages are obvious.

F. *Sources of losses*

Despite the early state of the art it is quite possible to get a general idea of the size of the losses which have to be expected in an electroluminescent *p–i–n* structure.

In order to keep the temperature of the device low to prevent thermal quenching of the luminescence, it has to be mounted on a heat sink just like a power diode. Ten watts per square centimeter can be safely dissipated for a small area without undue heating. If a temperature of 60°C above environment is tolerable, a good heat sink can dissipate 100 W/cm^2 in continuous operation. This sets the order of magnitude for all other values for our estimation.

The ohmic losses in the *i*-layer can be estimated both from the diffusion theory[58] and from the

space-charge-limited currents (SCLC) theory.[59] With assumptions similar to those made in the previous section, ohmic losses are below 10 per cent for the diffusion case. Higher losses will occur in the leads.

The SCLC theory in its present state gives only the ohmic losses, since recombination is assumed to be negligible and electrode and band-gap effects do not occur in this theory. The power set free by recombinations can be obtained from the current, assuming that each carrier creates a photon and that gains from the Peltier effect are negligible. The ohmic losses amount to 20–40 per cent of the total electrical input. Thus, the assumption of 50 per cent total loss due to joule heating is certainly a conservative one.

Forty per cent of the remaining 50 per cent of the input power is lost in radiationless transitions (assuming a quantum efficiency of 0·6 photons per current carrier; photoluminescent phosphors also have 60 per cent quantum efficiency). Of the remaining 30 per cent of the input power, 25 per cent is lost by the "quantum deficit" due to the energy difference between band-gap and impurity-center transition. If phosphors with edge emission could be used, this source of loss would be negligible. Of the remaining 22·5 per cent of the input power, which represents actually emitted light, 50–60 per cent is lost due to reabsorption before it is able to escape from the crystal. We have to assume this high loss figure because in most cases the anodic, hole-injecting contact will be opaque and will absorb most of the light falling on it. This illustrates the importance of finding new, transparent *p*-type materials suitable for hole emission. The reflection at the front surface can be reduced by using a Weierstrass sphere,[63] antireflex coatings, minute scattering bubbles in the crystals, or rough surfaces. The transparent, conducting layer at the cathode will also absorb 10–20 per cent which is contained in the above figure. The absorption of luminescence within the crystal itself has been neglected because it leads to the creation of new carrier pairs.

We end up with about 10 per cent efficiency for the total device, or 40 lm/W for an emission band which has an equivalent of 400 lm/W. This is very close to the maximum efficiency calculated for Destriau EL.[2] However, in the injection EL case there is much space left for improvements.

If 40 W/cm^2 of joule heat can be tolerated without excessive heating of the device, then we obtain a brightness of nearly $1{\cdot}5 \times 10^6$ ft-L for green light. Short pulses can be made much brighter. This is a decisive advantage of injection EL over Destriau EL, where 500 ft-L is regarded as extremely bright.

CONCLUSION

The mechanism of injection EL is much easier to understand than that of impact-ionization EL. Injection EL in junctions in single crystals offers high theoretical maximum values for efficiency and, especially brightness. A survey of promising host crystal materials is given. The discrepancies between the requirements for making luminescent *p–n* junctions and the properties of available materials are pointed out. The *p–i–n* concept offers more promise. The status of available theories is described. The practical difficulties of making ohmic contacts to wide-band-gap materials have been discussed. An estimation of sources of loss is given.

No theoretical or practical difficulties of insurmountable nature could be found which would render the pursuit of the concept of injection EL infeasible. However, the work of many people is still required to elucidate the details and to reduce the general idea to practice.

Acknowledgements—Gratitude is expressed to Mr. H. W. Leverenz, Drs. E. E. Loebner, R. E. Shrader, S. Larach and the Editors of this Journal for comments on the manuscript.

REFERENCES

1. For references see H. F. Ivey, *Trans. I.R.E.* ED-**6**, 203 (1959); S. Larach and R. E. Shrader, *RCA Rev.* 532 (1959); A. Fischer, *Z. angew. Phys.* **7**, 258 (1955).
2. F. Morehead, *J. Electrochem. Soc.* **107**, 281 (1960).
3. A. Fischer, *Conference Garmisch*, 1956, Semiconductors and Phosphors, p. 551. Fr. Vieweg, Braunschweig (1958).
4. E. A. Einstein, *Brit. J. Appl. Phys.* **8**, 190 (1957).
5. K. Lehovec, C. A. Accardo and E. Jamgochian, *Phys. Rev.* **83**, 603 (1951); **89**, 20 (1953).
6. J. Tauc, *Czech. J. Phys.* **7**, 275 (1957).
7. O. W. Lossev. See literature list of H. F. Ivey[1].
8. H. Tetzner, *Z. angew. Phys.* **1**, 153 (1948).
9. L. Patrick, *J. Appl. Phys.* **28**, 765 (1957); *International Solid State Conference, Brussels* (1958); W. J. Choyke and Lyle Patrick, *Phys. Rev.* **105**, 1721 (1957).

10. D. Rücker, *Z. angew. Phys.* **10,** 254 (1958).
11. A. Fischer, Thesis, University of Giessen, Germany (1957). (Partially unpublished.)
12. R. N. Hall, *J. Appl. Phys.* **29,** 914 (1958).
13. J. R. Haynes and H. B. Briggs, *Bull. Amer. Phys. Soc.* **2,** 14 (1952); *Phys. Rev.* **86,** 647 (1952); **98,** 1866 (1955); **101,** 1676 (1956); H. Briggs, *U.S. Pat.* 2683794 (1954).
14. R. Newman, *Phys. Rev.* **91,** 1313 (1953); **100,** 700 (1955); **105,** 1715 (1957).
15. J. S. Moss and J. H. Hawkins, *Phys. Rev.* **101,** 1609 (1956); *J. Phys. Radium* **17,** 712 (1956).
16. R. Braunstein, *Phys. Rev.* **99,** 1892 (1955).
17. C. Z. van Doren and D. de Nobel, *Physica, 's Grav.* **22,** 338 (1956).
18. E. E. Loebner and E. W. Poor, *Phys. Rev. Letters* **3,** 23 (1959).
19. G. J. Frosch and M. Gershenzon. Unpublished work.
20. M. S. Weinstein. To be published.
21. F. E. Williams, *E.C.S. Meeting, Chicago,* 1960. Enlarged Abstract No. 41.
22. K. Lehovec, *U.S. Pat.* 2776367 (1957).
23. E. E. Loebner, *U.S. Pat.* 2817783 (1957).
24. G. Diemer, *Electronics* **33,** 71 (1960).
25. A. Fischer, *Ger. Pat.* 1052563 (1960).
26. A. Fischer, *Z. Phys.* **149,** 107 (1957); R. Wolfe and J. Woods, *Phys. Rev.* **105,** 921 (1957); E. Krautz and G. Zollfrank, *Optik, Stuttgart* **14,** 446 (1957).
27. J. A. Lely, *Ber. dtsch. keram. Ges.* **32,** 229 (1955).
28. C. A. A. J. Greebe and W. F. Knippenberg, *Philips Res. Rep.* **15,** 120 (1960).
29. R. H. Wentorf, *J. Chem. Phys.* **26,** 956 (1957).
30. A. Fischer, *Ger. Pat.* 1057581 (1959).
31. B. Stone and D. Hill, *Phys. Rev. Letters* **4,** 282 (1960).
32. J. A. Perri, S. LaPlaca and B. Post, *Acta Cryst.* **11,** 310 (1958).
33. A. Fischer, *Ger. Pat.* 970869 (1958).
34. K. M. Taylor and C. Lenie, *J. Electrochem. Soc.* **107,** 308 (1960).
35. T. Renner, *Z. anorg. Chem.* **298,** 22 (1959).
36. H. G. Grimmeis, R. Groth and J. Maak, *Z. Naturf.* **15**a, 799 (1960); H. G. Grimmeis and H. Koelmans, *Ibid.* **14**a, 264 (1959).
37. C. J. Frosch, M. Gershenzon, D. F. Gibbs and R. M. Mykulyak, *E.C.S. Meeting, Chicago,* 1960. Electronics Division Abstracts, Vol. 9, No. 1, pp. 212, 215; also Personal communication (1960).
38. A. Fischer, *Ger. Pat.* 1074557 (1960).
39. J. W. Nielson and E. F. Dearborn, *J. Phys. Chem.* **64,** 1762 (1960).
40. J. Rudolph, *Z. Naturf.* **14**a, 727 (1959).
41. A. Fischer, *Z. Naturf.* **8**a, 756 (1953).
42. (a) A. Addamiano and P. A. Dell, *J. Phys. Chem.* **61,** 1020 (1957); (b) *Ibid.* **61,** 1253 (1957); (c) W. E. Medcalf and R. H. Fahrig, *J. Electrochem. Soc.* **105,** 719 (1958); (d) A. Addamiano and M. Aven, *J. Appl. Phys.* **31,** 36 (1960).
43. A. Fischer, *Z. Naturf.* **13**a, 105 (1958).
44. A. Fischer, *J. Electrochem. Soc.* **106,** 838 (1959).
45. A. Fischer, *Bull. Amer. Phys. Soc.* II **6,** 17 (1961); to be published (1961).
46. R. M. Potter, *E.C.S. Meeting, Chicago,* 1960. Enlarged Abstract, p. 79.
47. D. C. Reynolds, L. C. Greene, R. G. Wheeler and R. S. Hogan, *Bull. Amer. Phys. Soc.* II **1,** 111 (1956).
48. S. Larach, R. E. Shrader and C. F. Stocker, *Phys. Rev.* **108,** 587 (1957).
49. K. Baedecker, *Ann. Phys. Lpz.* **29,** 566 (1909).
50. J. S. Prener and F. E. Williams, *J. Phys. Chem. Solids* **8,** 461 (1959).
51. W. van Roosbroeck and W. Shockley, *Phys. Rev.* **94,** 1558 (1954).
52. L. Bess, *Phys. Rev.* **105,** 1469 (1957).
53. W. Shockley, *U.S. Pat.* 2569347 (1951); H. Kroemer, *Proc. I.R.E.* **45,** 1535 (1957); D. A. Jenny, *Ibid.* **46,** 959 (1958).
54. K. Lehovec, *Proc. I.R.E.* **40,** 1407 (1952).
55. A. Fischer, *Z. Naturf.* **9**a, 508 (1954).
56. E. R. Olson and E. H. Lougher, *Elec. Mfg.* **61,** 143 (1958).
57. Y. T. Sihvonen and D. R. Boyd, *Rev. Sci. Instrum.* **31,** 992 (1960).
58. R. N. Hall and W. C. Dunlap, Jr., *Phys. Rev.* **80,** 467 (1950); A. Herlet and E. Spenke, *Z. angew. Phys.* **7,** 99, 149, 195 (1955).
59. M. A. Lampert, *RCA Rev.* 682 (1959).
60. D. O. North. Personal communication (1960).
61. W. Lehmann, *J. Electrochem. Soc.* **107,** 657 (1960).
62. R. Mayrowitz, *Amer. Min.* **40,** 398 (1955).
63. P. Aigrain and C. Benoit-à-la-Guillaume, *J. Phys. Radium* **17,** 709 (1957).
64. A. Fischer, *Phys. Verh., Mosbach* **8,** 204 (1957).
65. W. Gordy and W. J. O. Thomas, *J. Chem. Phys.* **24,** 439 (1956).
66. F. N. Hooge, *Z. Phys. Chem.* **24,** 275 (1960).
67. H. G. Grimmeis, W. Kischio and A. Rabenau, *J. Phys. Chem. Solids* **16,** 302 (1960).
68. E. Loh and R. Newman, *Bull. Amer. Phys. Soc.* II **6,** 148 (1961).
69. R. A. Laudise and A. A. Ballman. To be published.
70. A. Fischer. Unpublished result.
71. W. Shockley and R. Hensley, *Bull. Amer. Phys. Soc.* II **6,** 106 (1961).
72. H. G. Grimmeis and H. Koelmans, *Philips Res. Rep.* **15,** 290 (1960).
73. N. A. Vitovskii and P. I. Maleev, *Soviet Phys., Solid State* **1,** 901 (1959).
74. R. H. Bube. Private communication.
75. R. W. Smith, *Phys. Rev.* **105,** 900 (1957).
76. R. H. Bube and E. L. Lind, *Phys. Rev.* **110,** 1040 (1958).
77. H. Koelmans, *J. Phys. Chem. Solids* **17,** 69 (1960).
78. R. W. Smith, *Phys. Rev.* **97,** 1525 (1955).
79. M. A. Lampert. To be published.

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 247–258. Printed in Great Britain

ALPHA CUTOFF FREQUENCY OF JUNCTION TRANSISTORS

J. LINDMAYER and C. WRIGLEY

Solid State Research and Development, Sprague Electric Company, North Adams, Mass.

(*Received* 28 *November* 1960; *in revised form* 6 *February* 1961)

Abstract—The alpha cutoff frequencies of drift and diffusion transistors are analyzed and approximated by simple analytical expressions. The alpha cutoff frequency is determined in terms of the three major contributions: base-region transport cutoff, emitter cutoff due to the emitter capacitance, and the collector cutoff due to the base resistance and the collector capacitance. Denoting the respective cutoff frequencies by $\omega_{c\alpha}^*$, ω_E and ω_c, and the resulting cutoff frequency by $\omega_{c\alpha}$, it is found by computation that

$$\frac{1}{\omega_{c\alpha}} = \frac{1}{\omega_{c\alpha}^*} + \frac{1}{\omega_E} + \frac{2}{\omega_c}$$

In addition, a new expression is given for the base transport cutoff.

The dependence of the frequency cutoff on the emitter current is also discussed for the current range where the built-in electric field is partially washed out by the high concentration of minority carriers injected into the base. A calculation is presented for the current level at which there is a maximum in the alpha cutoff.

Experimental data are compared with the calculations.

Résumé—Les limites de fréquence du facteur alpha des transistors de diffusion et d'apport sont analysées et on obtient des valeurs approximatives par de simples expressions analytiques. La limite de fréquence est déterminée en fonction des trois facteurs les plus importants: l'interruption de transport dans la région de base, l'interruption de l'émetteur due à la capacité de l'émetteur et l'interruption du collecteur due à la résistance de base et à la capacité du collecteur. En dénotant les limites de fréquence d'interruption respectives par $\omega_{c\alpha}^*$, ω_E et ω_c, et la fréquence d'interruption qui en résulte par $\omega_{c\alpha}$, on dérive l'expression suivante:

$$\frac{1}{\omega_{c\alpha}} = \frac{1}{\omega_{c\alpha}^*} + \frac{1}{\omega_E} + \frac{2}{\omega_c}$$

On dérive aussi une nouvelle expression pour la limite du transport de base.

La dépendence de la limite de fréquence en fonction du courant émetteur est aussi discutée pour la gamme de courant ou le champ électrique est partiellement emporté par la forte concentration de porteurs minoritaires injectés dans la base. On présente une expression pour le niveau du courant où un maximum de la limite de fréquence du facteur alpha existe. Les donneés expérimentales sont comparées avec les valeurs calculées.

Zusammenfassung—Die α-Grenzfrequenzen von Drift- und Diffusionstransistoren werden analysiert und in Annäherung durch einfache mathematische Ausdrücke dargestellt. Die α- Grenzfrequenz wird durch drei Hauptfaktoren bestimmt: die Grenzfrequenz für den Transport in der Basiszone, die von der Kapazität des Emitters abhängende Grenzfrequenz und die vom Basiswiderstand und der Kollektorkapazität abhängende Grenzfrequenz. Bezeichnen wir diese Frequenzen mit $\omega_{c\alpha}^*$, ω_E und ω_C und die resultierende Grenzfrequenz mit $\omega_{c\alpha}$, so ergibt sich durch Berechnung der Ausdruck:

$$\frac{1}{\omega_{c\alpha}} = \frac{1}{\omega_{c\alpha}^*} + \frac{1}{\omega_E} + \frac{2}{\omega_c}$$

Zusätzlich hierzu wird ein neuer Ausdruck für die Grenzfrequenz des Transports in der Basis gegeben.

Es folgt eine Erörterung der Abhängigkeit der Grenzfrequenz vom Emitterstrom innerhalb des Bereiches, in dem das eingebaute elektrische Feld durch die hohe Dichte der injizierten Minoritätsträger teilweise vernichtet wird. Das Stromniveau, für das die α-Grenzfrequenz ein Maximum erreicht, wird berechnet.

Die Berechnungen werden mit experimentellen Ergebnissen verglichen.

I. INTRODUCTION

THE high-frequency performance of early diffusion-type transistors was limited by the transit time of minority carriers diffusing across the base region.[1] This transit time decreases by a factor of 2 for high-injection levels.[2] Later, a gradient of impurities was incorporated into the base region in order to provide a built-in electric field[3] which increases the cutoff frequency.

However, there are other factors, such as depletion-layer capacitances and base resistance, which also affect the high-frequency operation of a transistor. Several aspects of the problem have been considered, such as development of an approximate expression for the intrinsic alpha cutoff[3], an analysis of depletion-layer cutoff frequency[4], and inclusion of the emitter depletion-layer capacitance[5]. The purpose of this paper is to provide manageable expressions for the alpha cutoff frequency as a function of the important structural parameters.

II. SYMBOLS AND UNITS IN ORDER OF APPEARANCE

E_{b0}	"built-in" field in the base region (V/cm)
k	Bolzmann constant (V-C/°K)
T	absolute temperature (°K)
q	electronic charge (C)
N	impurity concentration (cm^{-3})
N_E	impurity concentration at the emitter space-charge boundary (cm^{-3})
N_W	impurity concentration at the collector space-charge boundary (cm^{-3})
W	base width (cm)
x_0	length for $1/e$ fall of the exponential N (cm)
η	normalized "built-in" field (dimensionless)
α^*	complex intrinsic current transport factor in the base (dimensionless)
α_i^*	imaginary component of α^* (dimensionless)
D	diffusion constant of minority carriers (cm^2/sec)
ω	angular frequency (sec^{-1})
Ω	angular frequency normalized by W^2/D (dimensionless)
$\omega_{c\alpha}^*$	angular cutoff frequency at which $\lvert\alpha^*\rvert = 1/\sqrt{(2)}$ (sec^{-1})
$\Omega_{c\alpha}^*$	normalized form of $\omega_{c\alpha}^*$ (dimensionless)
$f_{c\alpha}^*$	cutoff frequency (sec^{-1})
Ψ	improvement factor (dimensionless)
τ	base-region transit time (sec)
Ω_τ	normalized inverse of τ (dimensionless)
C_E	emitter depletion-layer capacitance (F)
A_E	area of emitter junction (cm^2)
ϵ_0	dielectric constant of vacuum (C/V-cm)
ϵ	relative dielectric constant of the base material (dimensionless)
V_E	potential drop across emitter depletion layer (V)
V_0	emitter-junction built-in potential (V)
I_E	emitter current (A)
I_{E0}	emitter-junction reverse-saturation current (A)
α_E	complex current-transport factor with C_E included, $r_bC_c = 0$ (dimensionless)
y_{21}^*	forward transfer admittance of intrinsic transistor (Ω^{-1})
y_{11}^*	input admittance of intrinsic transistor (Ω^{-1})
y_{22}^*	output admittance of intrinsic transistor (Ω^{-1})
y_{12}^*	reverse transfer admittance of intrinsic transistor (Ω^{-1})
Δy^*	determinant y^*-matrix (Ω^{-2})
j	$\sqrt{(-1)}$
ω_E	angular-emitter cutoff frequency (sec^{-1})
Ω_E	normalized-emitter cutoff frequency (dimensionless)
ω_{CE}	angular cutoff frequency modified by C_E (sec^{-1})
Ω_{CE}	normalized ω_{CE} (dimensionless)

Vol. 2
1961

Γ ratio $\Omega_{CE}/\Omega^*_{c\alpha}$ (dimensionless)
r_b base resistance (Ω)
C_c collector depletion-layer capacitance (F)
α_c complex current-transport factor with r_bC_c included, $C_E = 0$ (dimensionless)
ω_c angular collector cutoff frequency $(1/r_bC_c)$ (sec^{-1})
ω_{cc} angular cutoff frequency modified by r_bC_c (sec^{-1})
Υ ratio $\omega_{cc}/\omega^*_{c\alpha}$ (dimensionless)
$\omega_{c\alpha}$ angular cutoff frequency of complex current-transport factor with C_E and r_bC_c included (sec^{-1})
η effective current-dependent η for $\omega^*_{c\alpha}$ (dimensionless)
$\bar{\eta}_\tau$ effective current dependent η for Ω_τ (dimensionless)
δ normalized current (dimensionless)
δ_{max} value of δ at which $\omega_{c\alpha}$ has a maximum (dimensionless)
ζ additive term describing the change in emitter resistance (dimensionless)
C_N normalized emitter capacitance (dimensionless)

III. DISCUSSION OF APLHA CUTOFF FREQUENCY

The alpha cutoff frequency of transistors can be analyzed utilizing the generally accepted equivalent circuit shown in Fig. 1. The box marked "Intrinsic" in the figure accounts for the transfer of injected minority carriers from emitter to collector. If we assume an exponential impurity distribution in this region, a field, E_{b0}, is obtained and has been shown to be independent of position for low injection levels.[4] Therefore, using $N = N_E \exp(-x/x_0)$,

$$E_{b0} = -\frac{kT}{q}\frac{N'}{N} = \frac{kT}{q}\ln\frac{N_E}{N_W} \tag{1}$$

where kT/q = thermal voltage, N = impurity concentration, N_E = impurity concentration at the emitter, N_W = impurity concentration at W and W = base width.

For convenience, let us define the quantity

$$\eta = \ln\frac{N_E}{N_W} = \frac{W}{x_0} \tag{2}$$

A. *Intrinsic transistor*

The base transport factor, α^*, (for the purposes of this paper we may set $\alpha^* = 1$ for low frequencies) may be written as[6]

$$\alpha^* = \frac{\exp(\eta/2)}{\cosh\sqrt{\left[\left(\frac{\eta}{2}\right)^2 + j\Omega\right]} + \left\{\left(\frac{\eta}{2}\right)\Big/\sqrt{\left[\left(\frac{\eta}{2}\right)^2 + j\Omega\right]}\right\}\sinh\sqrt{\left[\left(\frac{\eta}{2}\right)^2 + j\Omega\right]}} \tag{3}$$

where a normalized frequency

$$\Omega = \omega W^2/D \tag{4}$$

has been used and D is the diffusion constant of minority carriers. The asterisk refers to parameters of the intrinsic transistor.

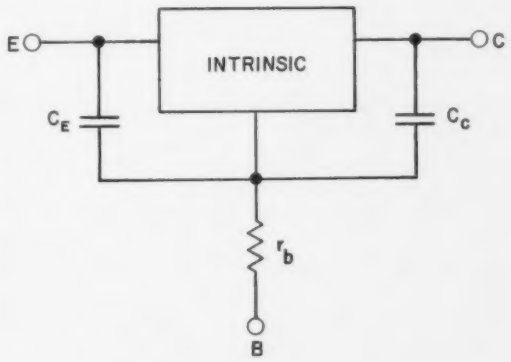


FIG. 1. Equivalent circuit for the transistor.

The complex current-transport factor, α^*, is plotted in Fig. 2. An approximate expression for the -3 db cutoff frequency of this intrinsic transport has been given by KRÖMER[3] for large η:

$$[\Omega^*_{c\alpha}]_{\text{KRÖMER}} \approx 2\left(\frac{\eta}{2}\right)^{3/2}$$

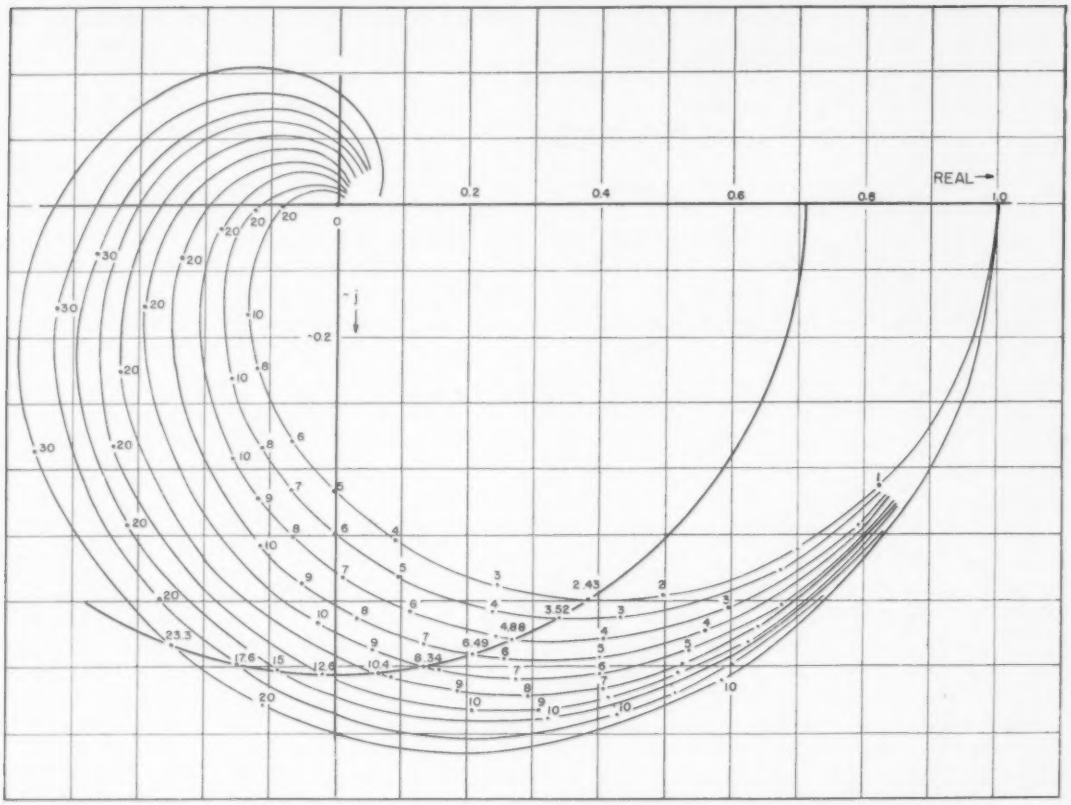


FIG. 2. The intrinsic-current transport factor, equation (3), plotted in the complex plane. The frequency scales along the curves are in units of W^2/D. The frequencies when the real part is 0·5 are the Ω_τ's. The alpha cutoff, $\Omega^*_{c\alpha}$, is the intersection of the transport factor with the $1/\sqrt{(2)}$ circle.

A computer was programmed to solve (3) for $\Omega = \Omega^*_{c\alpha}$ when $|\alpha^*| = 1/\sqrt{(2)}$. The resulting function is plotted in Fig. 3. It can be seen that KRÖMER's approximation is not very good for the range $\eta < 8$, which is used in actual drift transistors. A much better approximation is provided by the expression

$$\Omega^*_{c\alpha} = 2{\cdot}43\left[1+\left(\frac{\eta}{2}\right)^{4/3}\right] \qquad (5)$$

which is plotted in Fig. 3. It can be seen from the figure that (5) is a good approximation for $\eta \leqslant 10$, including zero.

The intrinsic alpha cutoff of a drift transistor is improved over that of a diffusion transistor by the factor

$$\psi = \frac{\Omega^*_{c\alpha}(\text{drift})}{\Omega^*_{c\alpha}(\text{diff.})} = 1+\left(\frac{\eta}{2}\right)^{4/3} \qquad (6)$$

Considering that the highest practical value of η is about 8, the improvement factor can be as large as 7·3. However, it will be shown that the improvement in the frequency cutoff of the entire transistor is a good deal less, owing to the effects of base resistance and depletion-layer capacitances.

It is interesting to compare the improvement of alpha cutoff with that of transit time. To the transit time, τ, there corresponds a frequency, Ω_τ, for which the real component of α^* in (3) equals 1/2. One has[7]:

$$\Omega\alpha = \frac{W^2}{D\tau} = \frac{\eta^2}{\eta-1+\exp(-\eta)} \qquad (7)$$

Vol. 2
1961

The transit time improves only by a factor of 4·6 as η increases from 0 to 8. The discrepancy between the η-dependence of $\Omega^*_{c\alpha}$ and Ω_τ is a consequence of the arbitrary -3 db definition for $\Omega^*_{c\alpha}$.

B. *Emitter efficiency*

The emitter depletion-layer capacitance is represented by C_E in Fig. 1. Considering that the emitter depletion layer is very narrow, we may neglect the change of impurity concentration within the depletion layer and have

$$C_E = A_E \sqrt{\left(\frac{N_E \epsilon \epsilon_0 q}{2V_E}\right)} \tag{8}$$

where A_E = emitter area, ϵ_0 = dielectric constant of vacuum, ϵ = relative dielectric constant of the base material, and

$$V_E = V_0 - \frac{kT}{q} \ln \frac{I_E}{I_{E0}}$$

is the potential drop across the emitter depletion layer, which includes the built-in potential V_0, a term depending logarithmically on the emitter current I_E and the emitter reverse-saturation current I_{E0}. Note that C_E varies little with I_E in the range of practical interest.

Since C_E is located at the input of the transistor, part of the a.c. emitter current will flow through C_E rather than be injected into the base as minority-carrier current. This is particularly important for drift transistors, where the high impurity concentration adjacent to the emitter results in a high emitter capacitance.

Let us next consider the hypothetical case where $r_b C_c$ is equal to 0, and calculate the cutoff frequency resulting from the emitter capacitance and from the intrinsic cutoff already discussed. The alpha vector, with C_E taken into account, can then be written as

$$\alpha_E = \frac{-y^*_{21}}{y^*_{11} + j\omega C_E} = \frac{\alpha^*}{1 + \dfrac{j\omega C_E}{y^*_{11}}} \tag{9}$$

where

$$y^*_{11} = \frac{qI_E}{kT} \frac{1-\exp(-\eta)}{\eta} \left\{ \frac{\eta}{2} + \sqrt{\left[\left(\frac{\eta}{2}\right)^2 + j\Omega\right]} \coth \sqrt{\left[\left(\frac{\eta}{2}\right)^2 + j\Omega\right]} \right\} \tag{10}$$

Equation (9) must now be solved for the normalized cutoff frequency $\Omega = \Omega_{CE}$ when $|\alpha_E| = 1/\sqrt{(2)}$. Utilizing (3), (4), (9) and (10) and introducing

$$\Omega_E = \frac{qI_E}{kT} \frac{1}{C_E} \frac{W^2}{D} \tag{11}$$

we obtain

$$|\alpha_E| = 1/\sqrt{(2)} = |\alpha^*|_{\Omega = \Omega_{CE}} = \frac{}{\left| (1+j)(\Omega_{CE}/\Omega_E) \Big/ \left\{ \frac{1-e^{-\eta}}{\eta} \left(\eta/2 + \sqrt{[(\eta/2)^2 + j\Omega_{CE}]} \coth \sqrt{[(\eta/2)^2 + j\Omega_{CE}]} \right) \right\} \right|} \tag{12}$$

The solution will provide the alpha cutoff frequency, Ω_{CE}, taking into account the base transport and emitter capacitance, but not collector capacitance and base resistance.

The function $\Omega_{CE} = f(\eta, \Omega_E)$ has been obtained by solving (12) numerically on a computer. In Fig. 4 we have plotted $\Gamma = \Omega_{CE}/\Omega^*_{c\alpha}$, [where $\Omega^*_{c\alpha}$ was computed using (2)], as a function of Ω_E with η as a parameter. As shown in Fig. 4, the results can be approximated with reasonable accuracy by

$$\Gamma = \frac{\Omega_{CE}}{\Omega^*_{c\alpha}} \simeq \frac{1}{1 + \dfrac{\Omega^*_{c\alpha}}{\Omega_E}}$$

$$\text{or} \quad \frac{1}{\omega_{CE}} \simeq \frac{1}{\omega^*_{c\alpha}} + \frac{1}{\omega_E} \tag{13}$$

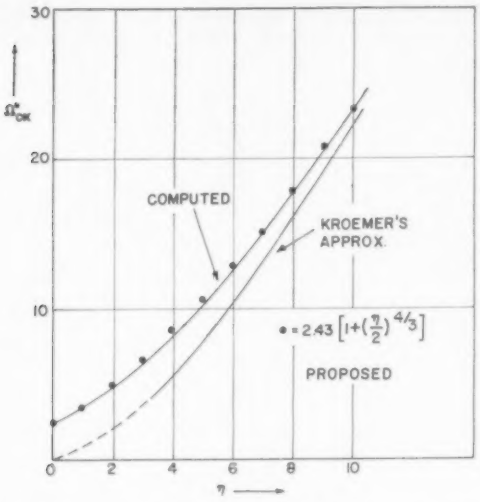


FIG. 3. Intrinsic cutoff frequency as a function of η: computed curve, KRÖMER's approximation and the new approximation (5).

Therefore, we have proved that reciprocal-frequency addition is reasonably accurate. This represents an approximation based on two *RC*-type cutoffs, but, since (2) is not exactly an *RC* transfer (Fig. 2), (13) is not exactly satisfied.

We can see from Fig. 4 that the high-frequency emitter efficiency Γ drops rapidly with increasing η. Fig. 5 has been obtained from the data of Fig. 4 by plotting Ω_{CE} as a function of η with Ω_E as a parameter. In order to hold Ω_E constant as η changes we can, for example, consider the values of N_E, W and I_E to be fixed, and change η by varying N_W.

As is seen from Fig. 5, the improvement of Ω_{CE} with increasing η is drastically reduced by the presence of Ω_E. As an example, using $W = 2\times10^{-4}$ cm, $D = 44$ cm^2/sec (p–n–p-Ge), $C_E = 30$ pF and $I_E = 5$ mA, we obtain the dashed curve in Fig. 5. Changing η from 0 to 8, the available improvement in cutoff frequency is only about a factor of 2, considerably less than the 7·3 obtained for the intrinsic transistor in (6).

Let us now examine the dependence of the alpha cutoff, Ω_{CE}, on I_E, in the low-current region. This dependence arises by means of the parameter Ω_E while $\Omega^*_{c\alpha}$ is independent of I_E.

Using the approximation (13), and changing from normalized frequencies to angular frequencies, we obtain

$$\frac{1}{\omega_{CE}} \simeq \frac{1}{\omega^*_{c\alpha}} + \frac{1}{I_E}\frac{kT}{q}C_E \qquad (14)$$

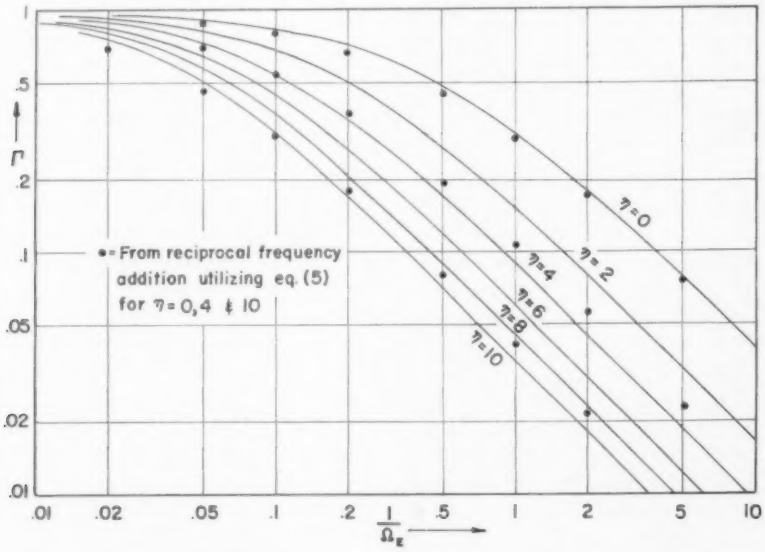


FIG. 4. The computed high-frequency emitter efficiency, Γ, and the approximation (13), as functions of $1/\Omega_E$ with η as a parameter.

Vol. 2
1961

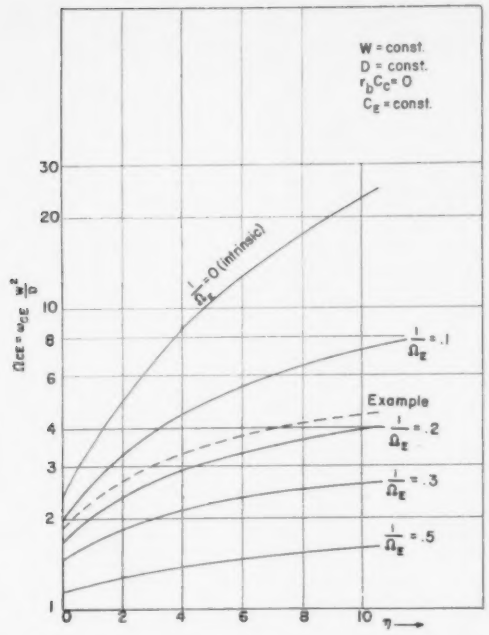


FIG. 5. The cutoff frequency with C_E included, Ω_{CE}, as a function of η with Ω_E as a parameter. The dashed curve is an example of a p–n–p germanium transistor with $D = 44$ cm²/sec, $W = 2\ \mu$, $C_E = 30$ pF and $I_E = 5$ mA.

Therefore,

$$\frac{d(1/\omega_{CE})}{d(1/I_E)} = \frac{kT}{q}C_E \tag{15}$$

Thus, for low injection, the rate of alpha cutoff improvement depends solely on C_E. Equation (15) will be discussed in more detail later.

C. *Collector cutoff*

Let us now consider the hypothetical case $C_E = 0$ and consider the effect of the base resistance r_b and the collector depletion layer capacitance C_c on the intrinsic transistor. The resulting alpha transport factor α_c may be calculated from Fig. 1, (with $C_E = 0$) as

$$\alpha_c = \frac{-y_{21}^* + r_b(\Delta y^* + y_{11}^* j\omega C_c)}{y_{11}^* + r_b(\Delta y^* + y_{11}^* j\omega C_c)} \tag{16}$$

where

$$\Delta y^* = y_{11}^* y_{22}^* - y_{21}^* y_{12}^*$$

Now, since

$$\alpha^* = -y_{21}^*/y_{11}^*$$

$$\alpha_c = \frac{\alpha^* + r_b(y_{22}^* + \alpha^* y_{12}^* + j\omega C_c)}{1 + r_b(y_{22}^* + \alpha^* y_{12}^* + j\omega C_c)} \tag{17}$$

In all practical cases y_{22}^* and y_{12}^* are quite small and we may approximate

$$\alpha_c = \frac{\alpha^* + j\omega r_b C_c}{1 + j\omega r_b C_c} = \frac{\alpha^* + j(\omega/\omega_c)}{1 + j(\omega/\omega_c)} \tag{18}$$

where $\omega_c = 1/r_b C_c$.

Solving (18) with $|\alpha_c| = 1/\sqrt{(2)}$, we obtain the cutoff frequency ω_{cc}.

$$\frac{\omega_{cc}}{\omega_c} = -2\alpha_i^* \pm \sqrt{(4\alpha_i^{*2} + 1 - 2|\alpha^*|^2)} \tag{19}$$

where α_i^* = imaginary component of α^* for $\omega = \omega_{cc}$ and $|\alpha^*|$ = absolute value of α^* for $\omega = \omega_{cc}$.

Since $\omega_{c\alpha}^* > \omega_{cc}$, the α^* vector at $\omega = \omega_{cc}$ has a smaller phase angle than at $\omega = \omega_{c\alpha}^*$. As is apparent from Fig. 2, α_i^* must then be negative for all practical values of η and both solutions of (19) will be positive.

The question of which solution is valid is resolved by reference to Fig. 6. The lowest curve is

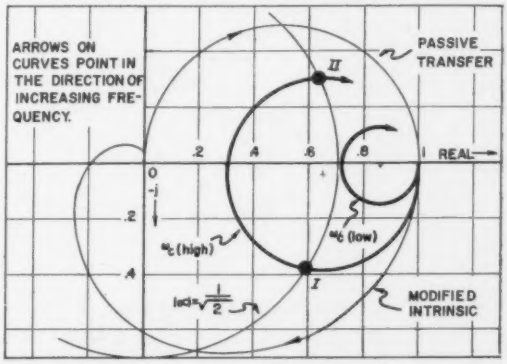


FIG. 6. Typical current transport factors, equation (18), plotted in the complex plane with $C_E = 0$, and $r_b C_c \neq 0$. The lowest curve is a modified intrinsic transport factor, and the semicircle in the upper quadrant is the passive transfer. The curves labeled "$\omega_{c(\text{high})}$" and "$\omega_{c(\text{low})}$" are the resulting α_c current-transport factors.

the locus of an intrinsic transfer vector α^*, modified by the factor $[1+j(\omega/\omega_c)]^{-1}$. This is the first term in (18). The top curve represents the passive transfer (majority current)

$$\frac{j(\omega/\omega_c)}{1+j(\omega/\omega_c)}$$

which is the second term in (18). Adding the two vectors at each ω results in the type of α_e locus represented by the heavy curves in Fig. 6. The two heavy curves shown are for two values of ω_c. If ω_c is too low (as for the curve marked $\omega_{c(\text{low})}$) the resulting α_e will never cross the $1\sqrt{(2)}$ circle, i.e. a -3 db cutoff is never reached and the square root becomes imaginary in (19). This occurs as ω_{cc} is dominated by ω_c; i.e. r_bC_c is so high that the passive transfer is dominant. However, if r_bC_c is so high, the transistor is so poor that it should not be considered.

For a reasonably low r_bC_c, (as for the curve marked $\omega_{c(\text{high})}$) the α_e locus has two intersections with the cutoff circle, I and II. Apparently intersection I is the one of interest, since it occurs as $|\alpha_e|$ is decreasing, and II occurs as $|\alpha_e|$ rises when the transistor becomes a passive network. Since intersection I is required, the negative sign applies in (19).

We now need a relationship linking, $\omega^*_{c\alpha}$, ω_c and ω_{cc}. For reasons we shall discuss later, it is convenient to compute and plot $\omega_{cc}/\omega^*_{c\alpha}$ vs. ω_{cc}/ω_c. The abscissa, the left-hand side of (19), is determined by a picked α^*, which also gives η.† From the η and the function represented in Fig. 3 we obtain $\omega_{cc}/\omega^*_{c\alpha}$. This ratio was computed for numerous α^* vectors and the result is plotted in Fig. 7. The ordinate may be considered as the high-frequency collector efficiency

$$\Upsilon = \frac{\omega_{cc}}{\omega^*_{c\alpha}} \tag{20}$$

A complete analytical expression for Υ would be complicated. Attempting to add the reciprocal frequencies $1/\omega^*_{c\alpha}$ and $1/\omega_c$ we find a poor approximation as the dotted curve in Fig. 7. However a

† The restrictions are: (a) $|\alpha^*| > 1/\sqrt{(2)}$ so that ω_c is finite and positive; (b) $|\alpha^*|^2 < \frac{1}{2}+2\alpha_i^{*2}$ so that ω_c is larger than ω_{cc}.

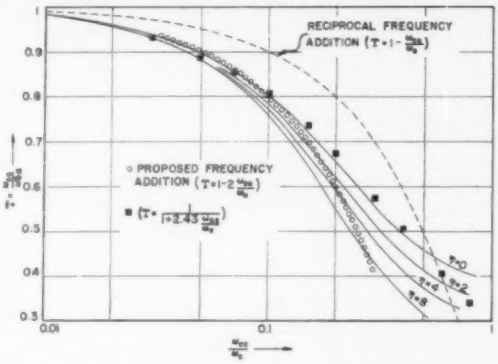


FIG. 7. The computed high-frequency collector efficiency Υ and three approximations. The squares are only for $\eta = 0$.

moderately good fit is obtained if $1/\omega_c$ is weighted by a factor of 2.

$$\frac{1}{\omega_{cc}} = \frac{1}{\omega^*_{c\alpha}} + \frac{2}{\omega_c} \text{ or } \Upsilon = 1-2\frac{\omega_{cc}}{\omega_c} \tag{21}$$

This approximation is plotted as the circles in Fig. 7. The approximation is fairly good for $\omega_c \geqslant 3\omega_{cc}$. A reasonably designed drift transistor will easily fulfill this requirement.

It is now seen that the plot in Fig. 7 is a convenient representation, since the ordinate is the high-frequency collector efficiency and the function only has a mild η dependence, suggesting the η independent approximation (21). A better approximation than (21) for the diffusion transistor would be

$$\Upsilon_{(\text{diff})} = \frac{1}{1+2{\cdot}43\,\omega_{cc}/\omega_c} \tag{22}$$

However, we will use (21) for all cases, since we are mainly concerned with drift transistors.

D. *Over-all cutoff with* Γ *and* Υ

We have considered emitter and collector efficiencies independently so far. It now remains to combine the two in order to obtain the over-all cutoff frequency. Since reciprocal frequency addition applies to the emitter it can be concluded that the general shape of the complex α^* function is not changed by addition of C_E, and the frequency

Vol. 2
1961

scale along the α_E curve is similar to the α^* scale.

The α^* frequency scale is dependent on η (Fig. 2), but Υ is fairly independent of η. Therefore Υ must be fairly independent of the magnitude of the frequency.

As a consequence of this argument, we may write (20) as

$$\Upsilon = \frac{\omega_{cc}}{\omega^*_{c\alpha}} \to \frac{\omega_{c\alpha}}{\omega_{CE}} \tag{23}$$

where $\omega_{c\alpha}$ is the over-all cutoff frequency.

Using (13) and (23) we may write the over-all efficiency as

$$\frac{\omega_{c\alpha}}{\omega^*_{c\iota}} = \Gamma\Upsilon \tag{24}$$

or

$$\frac{1}{\omega_{c\alpha}} = \frac{1}{\omega^*_{c\iota}} + \frac{1}{\omega_E} + \frac{2}{\omega_c} \tag{25}$$

By means of numerous assumptions we have now shown that the reciprocal over-all cutoff frequency, $1/\omega_{c\alpha}$, can be obtained by adding three inverse frequencies. The time constant, $1/\omega_c$, representing r_bC_c enters with a weight factor of 2, and the time constant $1/\omega^*_{c\alpha}$ should be obtained from (5).

IV. HIGH-CURRENT OPERATION

The only current-dependent term in (25) for low currents is $1/\omega_E$. For high current densities the current dependence of ω_E changes and $\omega^*_{c\alpha}$ also becomes current dependent. For discussion of the high-current case the following assumptions will be made: firstly, that the reciprocal-frequency addition in (25) holds for all currents, secondly, that ω_c is fairly current independent, which ties in with the third assumption that the base width is current independent. This is almost the case for abrupt collector junctions.

At high current densities the minority-carrier concentration in the base region becomes comparable in magnitude to the impurity concentration. The electric field in the base region is then dependent on the sum of the impurity and minority-carrier distributions and their slopes. For $\eta = 0$, RITTNER has shown[2] that $\Omega^*_{c\alpha}$ doubles as the current is increased indefinitely. This is equivalent to saying that η changes from 0 to 2 in (5).

For drift transistors too, as the current is increased, the minority carriers begin to dominate the base region, and the impurity distribution becomes unimportant. Therefore, all intrinsic transistors become identical with an effective $\eta = 2$, as the current is increased indefinitely. Therefore, it is convenient† to introduce an effective η which is current dependent: $\bar{\eta}$. Since the current dependence of $1/\Omega_\tau$ is known[7], and η is related to Ω_τ by (7), an effective $\bar{\eta}_\tau$ describing the current dependence of Ω_τ can be accurately constructed. This current-dependent $\bar{\eta}_\tau$ changes from $\bar{\eta}_\tau = \eta$ to $\bar{\eta}_\tau \simeq 2{\cdot}6$ as the current is increased. However, the $\bar{\eta}$ describing the current dependence of $\Omega^*_{c\alpha}$ must have the high-current limit $\bar{\eta} = 2$. Let us now obtain $\bar{\eta}$ as a function of current by assuming that $\bar{\eta}$ approaches its high-current limit to the same degree at a particular current as $\bar{\eta}_\tau$ approaches its own limit, i.e.

$$\bar{\eta} = 2+(\eta-2)(\bar{\eta}\tau-2{\cdot}6)/(\eta-2{\cdot}6)$$

This $\bar{\eta}$ is plotted in Fig. 8 as a function of current with the current-independent physical η as a parameter. Since the plot has only a few values of the parameter η it may be useful to have an approximate analytical form for $\bar{\eta}$. For $2 \leqslant \eta \leqslant 8$ (which is the practical range for drift transistors) the following expression fits within 10 per cent accuracy

$$\eta = \frac{\eta-2}{1+0{\cdot}735\delta\eta\exp(0{\cdot}03\eta)} + 2 \tag{26}$$

where

$$\delta = \frac{I_E W}{N_E q A_E D}$$

Now that we have the current dependence of $\omega^*_{c\alpha}$ in (25) let us consider the current dependence of ω_E. For low current densities ω_E depends linearly on I_E (11). However, for high current densities the input resistance kT/qI_E changes to

† If this approach were not taken it would be necessary to consider a non-linear partial differential equation, which has no general solution, in order to obtain the effects of high-current operation. The approach taken here results in reasonable agreement with empirical results.

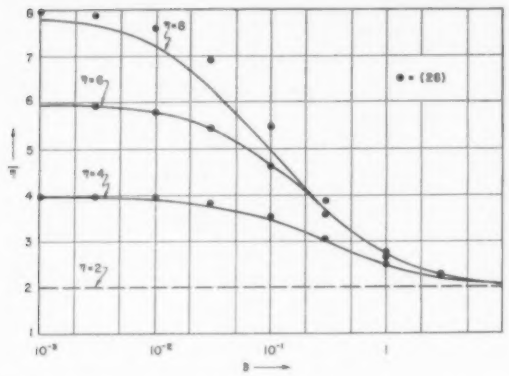


FIG. 8. The effective drift parameter $\bar{\eta}$ to be used in (5) for high current densities, as a function of δ with η as a parameter.

$2kT/qI_E$. This can be represented by

$$\text{Input resistance} = \frac{kT}{qI_E}(1+\zeta) \qquad (27)$$

where ζ changes from 0 to 1 as the current is increased. ζ has been computed[7] and is shown in Fig. 9. For current densities up to $\delta = 0{\cdot}2$ and for drift transistors ($\eta \geqslant 2$), the following simple form is applicable:

$$\zeta = \frac{\delta}{\eta} \qquad (28)$$

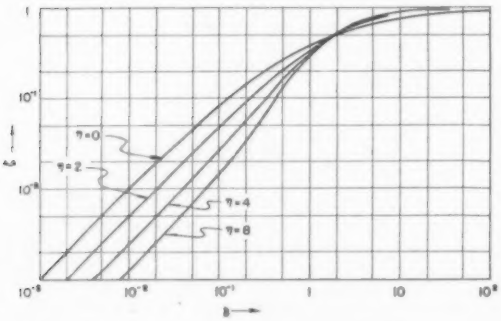


FIG. 9. The additive term ζ modifying the differential input resistance as a function of δ, with η as a parameter.

Either this form or the plot of ζ can be used to obtain ω_E.

From determination of the current dependence of the first two terms of the right-hand side of (25) and assumption of a current-independent third term, the current dependence of $\omega_{c\alpha}$ is obtained. Since $1/\omega_E$ is decreasing with increasing current density, and $1/\omega_{c\alpha}^*$ is increasing, a minimum in $1/\omega_{c\alpha}$ is expected at some current density. Using (26) and (27) in (25) one can find the normalized current $\delta_{\max}$, at which the maximum alpha cutoff occurs. The result is plotted in Fig. 10, where

$$C_N = \frac{kT}{q}\frac{C_E}{N_E q A_E W}$$

For $C_N > 10^{-2}$ the maximum tends to disappear, since the improvement of the emitter time constant with increasing current density is too slow

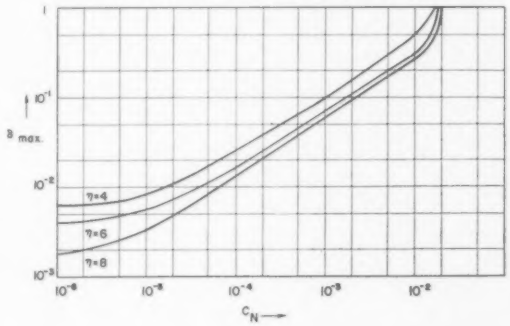


FIG. 10. The value $\delta = \delta_{\max}$, at which the over-all cutoff frequency has a maximum, as a function of C_N with η as a parameter.

compared to the deterioration of the intrinsic transfer function. The range $2\times10^{-5} < C_N < 5\times10^{-3}$ is typical for drift transistors, and in this range

$$\delta_{\max} \simeq \frac{44(C_N)^{2/3}}{\eta} \quad (\eta > 2) \qquad (29)$$

V. COMPARISON TO MEASUREMENT

In order to obtain the analytical approximation discussed in this paper, numerous simplifying assumptions had to be made. Therefore, it is desirable to compare the results with actual measurements. Fig. 11 shows an example of a p–n–p-Ge transistor with abrupt junctions. The theoretical

Vol. 2
1961

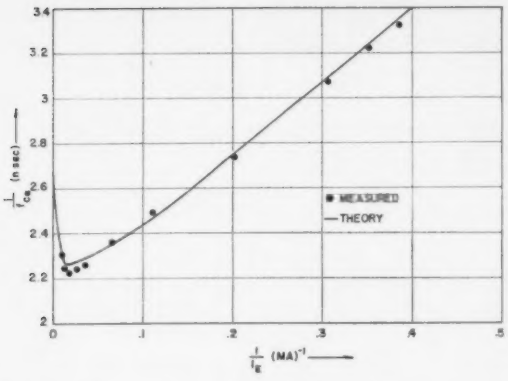


FIG. 11. Measured over-all cutoff frequency vs. current for a p–n–p germanium transistor.

curve was obtained from (25) using the measured values:

$$\eta = 4{\cdot}6$$
$$W = 2\times 10^{-4}\ \text{cm}$$
$$C_E = 20\ \text{pF}$$
$$r_bC_c = 120\ \text{psec}$$
$$A_E = 1{\cdot}27\times 10^{-4}\ \text{cm}^2$$
$$N_E = 10^{17}\ \text{cm}^{-3}$$
$$D = 44\ \text{cm}^2/\text{sec}$$

The points plotted in Fig. 11 were obtained by direct measurement. The agreement shows that the approximations on which (25) was based were justified. For this transistor we would expect the minimum $1/f_{c\alpha}$, from (29), to occur at $I_E = 72$ mA. The measurement shows a minimum at 67 mA. This is an error of $+7{\cdot}5$ per cent in the calculation. The maximum deviation of the calculated $1/f_{c\alpha}$ in Fig. 11 is 2 per cent. The expected low-current slope, as given in (15), is indistinguishable from measurement.

Extrapolation of the low-current slope to infinite current ($1/I_E = 0$) supplies ω_{cc}. In this example $\omega_{cc}/2\pi = 472$ Mc/s. For this extrapolated case (25) can be written as

$$\frac{1}{\omega_{cc}} = \frac{1}{\omega^*_{c\alpha}} + \frac{2}{\omega_c}$$

Using the values obtained for ω_{cc} and ω_c, we obtain $f^*_{c\alpha} = 1{\cdot}63$ kMc/s. Using the measured values of η and W in (5) we obtain $f^*_{c\alpha} = 1{\cdot}71$ kMc/s.; i.e. agreement within 5 per cent.

VI. SUMMARY AND CONCLUSIONS

The objective of this paper was to find simple, manageable expressions for the alpha cutoff frequency of junction transistors, with particular emphasis on drift transistors. The approximation for the intrinsic cutoff frequency given by KRÖMER was found to be inaccurate for the range of practical drift transistors. A simple and accurate approximation was proposed and compared to the actual function obtained by computation. By computing the effect of the emitter depletion-layer capacitance, it was found that reciprocal-frequency addition is valid; within reasonable accuracy. The effect of the collector time constant was also computed and the reciprocal-frequency addition could be extended to the collector, using a weight factor of 2. The over-all performance could be evaluated by three reciprocal frequencies, taking into account the factor of 2 for r_bC_c.

Expressions were given for high-current operation, where the intrinsic cutoff becomes current dependent and the emitter cutoff is no longer linearly dependent on the current. The collector time constant was assumed to be current independent. The current density at which the over-all alpha cutoff reaches a maximum was calculated.

We have considered the major effects and neglected only the minor effects such as d.c. injection efficiency, recombination, current-crowding effect, etc. The good agreement obtained with measurement justifies the neglections.

The simple expressions for the over-all performance allow one to obtain the parameters important to high-frequency operation. These parameters are available from one graph, i.e. a plot of $1/\omega_{c\alpha}$ vs. $1/I_E$, which is the most useful form for plotting the current dependence of $\omega_{c\alpha}$. The depletion-layer capacitance can be calculated from the low-current slope. Extrapolation of this slope to the $1/\omega_{c\alpha}$ axis reveals the value of $1/\omega_{cc}$ at the intercept. By means of an independent measurement of r_bC_c the value of $\omega^*_{c\alpha}$ can be obtained. If either W or η is known, the other can be got from $\omega^*_{c\alpha}$. Furthermore, if either N_E or A_E is known, the other can be obtained from a

measurement of the current at which $\omega_{c\alpha}$ has a maximum.

Utilizing the expressions derived for the high-frequency performance, it is interesting to consider the effect of reducing all physical dimensions of the transistor by the same factor. Let us consider (25) term by term. The intrinsic term is proportional to the square of the physical size, and even approximately so at δ_{max}. The emitter term is also proportional to the square of the size due to the emitter area. However, $r_b C_e$ appears to remain constant since r_b is about inversely proportional to the size whereas C_e is directly proportional to the size (the collector space-charge depth is also reduced, by reduction of the collector voltage). Consequently, a size reduction leads to an alpha cutoff dominated by $r_b C_e$. This points out the necessity of reducing $r_b C_e$ by some independent means if size reduction is to be employed. Otherwise

$$\frac{1}{\omega_{c\alpha}} \simeq \text{const}_1\,(\text{size})^2 + \text{const}_2$$

Knowing the expression for the current at the point of maximum performance (29), we can estimate the d.c. power required to stay at this point as the size is reduced. From (29) and the definition of δ, it can be seen that the current at the maximum depends on the cube root of the size. The collector voltage necessary to maintain a fixed ratio between base width and collector space-charge width is porportional to the square of the size. Therefore, the necessary d.c. power is

$$\text{Power} \simeq \text{const}_3\,(\text{size})^{7/3}$$

Thus, the power per unit collector area will be

$$\frac{\text{Power}}{\text{Collector area}} \simeq \text{Const}_4\,(\text{size})^{1/3}$$

The power dissipation per unit area decreases slowly, with the cube root of the size reduction. However, this holds only as long as $C_N < 10^{-2}$.

A decrease in physical size results in some improvement in alpha cutoff and power dissipation per unit area. However, r_b should be decreased by some independent method.

Acknowledgements—The writers are grateful to K. LEHOVEC for helpful discussion and advice, and to K. SCHOENI for the computations.

REFERENCES

1. W. SHOCKLEY, *Bell Syst. Tech. J.* **28,** 435 (1949).
2. E. S. RITTNER, *Phys. Rev.* **94,** 1171 (1954).
3. H. KRÖMER, *Arch. elektr. Übertr.* **8,** 364 (1954).
4. H. KRÖMER, The drift transistor, *Transistors I.* RCA Labs, Princeton (1956).
5. A. L. KESTENBAUM and N. H. DITRICK, *RCA Rev.* **18,** 17 (1957).
6. D. THOMAS and J. MOLL, *Proc. I.R.E.* **46,** 1178 (1958).
7. J. LINDMAYER and C. WRIGLEY, *Solid-State Electron.* **2,** 79 (1961).

Solid-State Electronics Pergamon Press 1961. Vol. 2, pp. 259–261. Printed in Great Britain

NOTES

Superconducting vanadium thin films*

(*Received* 16 *February* 1961)

THE study of the properties of vanadium thin films has been extended from the work previously reported.[1] Recent results in this continuing study have resulted in the observation of superconductivity in vanadium thin films. Two specimens gave quite different results in the resistance versus temperature behavior (see Fig. 1).

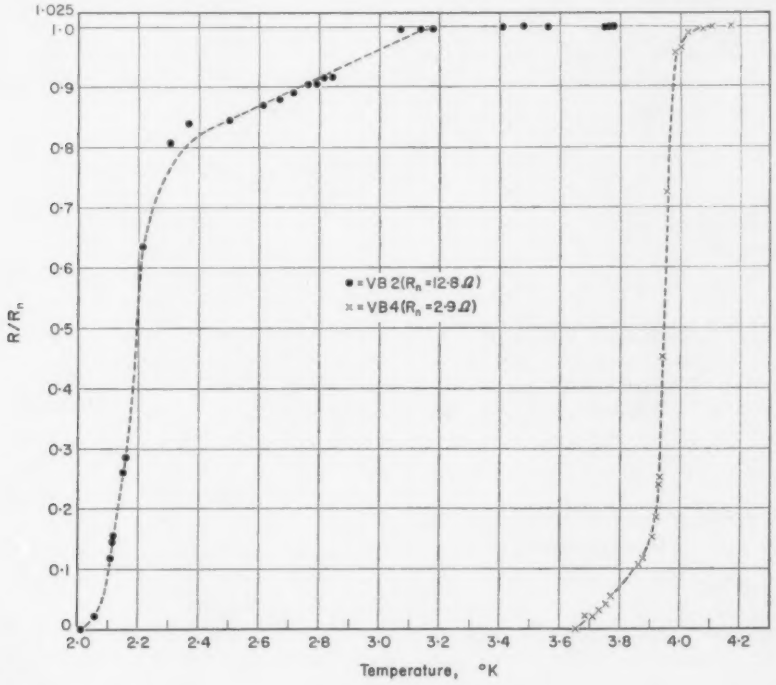


FIG. 1.

One specimen, designated VB2, has a rather broad transition beginning at 3·18°K and continuing to zero resistance at 2·02°K. There is no discontinuous transition observed. This sample was deposited between 4×10^{-7} and 9×10^{-7} mm Hg in 3 min. The source material had been outgassed at 1500°C and at 5×10^{-6} mm Hg for 24 hr prior to deposition. The substrate was a glass micro-slide. The dimensions of the film were 1·0 mm in width, 15·0 mm in length and on the order of 1350 Å in thickness.

Another specimen, designated VB4, has a resistance–temperature curve which shows a transition beginning at 4·10°K and continuing to zero resistance at 3·66°K. The transition is much sharper than in sample VB2. The film VB4 was of the same dimensions as VB2. The deposition was carried out in 4½ min at a pressure between 4×10^{-7} and 6×10^{-7} mm Hg. The source was

* This work supported in part by the Office of Naval Research, contract NONR 2145(00).

thoroughly outgassed having been melted at low pressures during three previous depositions and having been baked for 24 hr at 5×10^{-6} mm Hg prior to deposition. The substrate of VB4 was quartz.

The ratio of resistance at 4·2°K and at 295°K was 0·30 for the latter specimen and 0·53 for the former. It is believed that the gaseous impurity content of VB4 is less than that of VB2 and that this is the main cause of the difference in the transition curves.

Mechanically the films are very tough. The adherence of the vanadium films is extraordinary, there being no visible film damage after drawing a steel scriber across the film. This strong mechanical property has also been reported for tantalum films by MARCHAND and VENEMA[2].

Further work is being done on depositions at lower pressures. The effects of magnetic fields on the transitions are also being studied and will be reported in detail at a later date.

C. T. THOMPSON
J. F. GERBER

Burroughs Laboratories
Burroughs Corporation
Paoli, Pennsylvania

REFERENCES

1. J. D. BLADES, J. F. GERBER and C. T. THOMPSON, *Solid-State Electron.* **1**, 319 (1960).
2. J. F. MARCHAND and A. VENEMA, *Philips Res. Rep.* **14**, 427 (1959).

Distribution of thermal-conversion centers in germanium with indiffused impurities

(*Received* 1 *March* 1961)

WE SHOULD like to comment on a possible dependence of thermal conversion in the bulk of germanium samples on indiffusion of impurities of the third and fifth column of the periodic table into the surface of the sample. We have found some experimental evidence for such a relation.

FRANK and TURNBULL[1] and VAN DER MAESEN and BRENKMANN[2] have considered the equilibrium between vacancies, interstitial copper and substitutional copper in germanium to explain certain observations regarding thermal-conversion centers. Substitutional copper is believed to constitute the thermal-conversion centers; however, its motion results from dissociation into a fast-moving interstitial copper and a vacancy, with subsequent recombination to form again substitutional copper.

It is well known[3] that certain impurities, e.g. three- or five-valent elements diffusing into germanium, cause an electric potential to arise across the indiffused layer during the diffusion, which is required to balance the flow rates of electronic carriers and of indiffusing impurities. Considering that the mobility of electronic carriers is very large compared to that of the diffusing impurities, one has

$$V_S = \pm \frac{kT}{q} \ln \frac{C_S}{n_i} \qquad (1)$$

where kT/q is the voltage equivalent of temperature, C_S is the surface concentration of impurities and n_i is the intrinsic carrier concentration at the temperature of diffusion. Equation (1) applies if $C_S \gg n_i$, and if the bulk doping level is small compared to n_i. The surface becomes positively charged versus the bulk in the case of diffusing donor-type impurities, and negatively charged in the case of acceptor-type impurities.

Since vacancies are acceptors and have, therefore, a net negative charge and since they are much more mobile than the indiffusing impurities, it follows that their concentrations in the bulk and at the surface should differ by the Boltzmann factor:

$$\exp(+qV_S/kT) = (C_S/n_i)^{\pm 1} \qquad (2)$$

The sign of the factor is such that indiffusing n-type impurities suppress the concentration of vacancies in the bulk with respect to that at the surface, while for indiffusing p-type impurities the bulk vacancy concentration is enhanced with respect to that at the surface. The surface concentration of vacancies should be independent of the indiffusion of impurities except in the case that the free energy for formation of a Schottky defect is affected by interaction with impurities present in the crystal lattice.

Thermal conversion in copper-containing samples is usually attributed to substitutional copper rather than to vacancies. The distribution

Vol. 2
1961

of substitutional copper is determined by both vacancies and interstitial copper. The distribution of interstitial copper depends on its effective charge.

If interstitial copper were electrically neutral, a homogeneous distribution in the electric field adjacent to the surface should be expected and the distribution of substitutional copper should be a Boltzmann function, like that of the vacancies. In this case it is evident that indiffusion of *n*-type impurities suppresses bulk conversion, while indiffusion of *p*-type impurities enhances bulk conversion comparing samples of equal copper content.

In the case that interstitial copper were donors, as is perhaps likely, the Boltzmann distributions of interstitial copper and of vacancies would be reciprocal to each other, and their product, which is proportional to the concentration of substitutional copper, would become space independent. However, an influence of the indiffused impurities on the bulk thermal-conversion level may be expected from the following argument: *n*-type indiffused impurities cause a field which promotes a high surface concentration of interstitial donor-type copper (with respect to the bulk) and thus enhances the evaporation rate of copper from the sample. *p*-Type indiffused impurities promote a low surface concentration of interstitial donor-type copper and thus retard its evaporation rate from the sample.

In the course of studying diffusion of P and As into germanium of low-bulk impurity concentrations ($\sim 10^{14}$ cm^{-3}) we have observed frequent thermal conversion to occur at low surface concentrations, but never at surface concentrations larger than 5×10^{18} cm^{-3} in the case of As, and $1{\cdot}2 \times 10^{18}$ cm^{-3} in the case of phosphorus.[4]

Thus we have found it expedient to prevent thermal conversion of the bulk of the sample, when investigating a diffusion cycle of low surface concentration, by adding a shorter diffusion cycle of a sufficiently high surface concentration. This short added diffusion cycle does not modify significantly the distribution of impurities arising from the main cycle except near the surface, and did not interfere, therefore, with the evaluation method used by us for the impurity distribution resulting from the main diffusion cycle.

Unpublished work by BUSEN and MEEKS[5] on the diffusion of indium into germanium suggests that the thermal-conversion level of the bulk decreases toward the surface of the sample in qualitative agreement with an interpretation advanced in this Note.

KURT LEHOVEC

Sprague Electric Company
North Adams
Massachusetts

REFERENCES

1. F. C. FRANK and D. TURNBULL, *Phys. Rev.* **104**, 617 (1956).
2. F. VAN DER MAESEN and J. A. BRENKMANN, *J. Electrochem. Soc.* **102**, 229 (1955).
3. F. M. SMITS, *Proc. I.R.E.* **46**, 1049 (1958).
4. K. LEHOVEC, C. PIHL and C. WRIGLEY, *Gaseous Diffusion of Phosphorus in Germanium*, 115th meeting of Electrochemical Society, May, 1959; K. LEHOVEC and C. PIHL, *J. Electrochem. Soc.* To be published (1961).
5. K. BUSEN and E. MEEKS, *Surface Concentration of Indium in Germanium from Gaseous Diffusion, II.* Presented at the 117th meeting of the Electrochemical Society, May, 1960.

Papers to be published in future issues

M. GREEN, I. N. GREENBERG, J. BANZHAF, A. SASS, H. S. WHITE and R. V. MCKNIGHT: Electrical characteristics of a reduced-area, rhodium-plated, silicon end contact

W. T. ERIKSEN and H. AHLBURG: An analysis of electroluminescent brightness waves

C. FA and R. ZULEEG: Field-retarded diffusion of antimony in germanium

W. HEYWANG: Bariumtitanat als sperrschichthalbleiter

D. F. EDWARDS, C. D. BUTTER and L. D. MCGLAUCHLIN: Photoconductivity in single-crystal tellurium

W. RINDNER: A negative-resistance photoelectric diode

D. T. J. HURLE: Constitutional supercooling during crystal growth from stirred melts—I. Theoretical

K. SATO and M. TOMONO: Current dependence of α_{cb} in the alloy junction transistor

K. LEHOVEC and A. SLOBODSKOY: Diffusion of charged particles into a semiconductor under consideration of the built-in field

WEN-HSIUNG KO: The forward transient behavior of semiconductor junction diodes

J. E. PARROTT and A. W. PENN: The design theory of thermoelectric cooling elements and units

F. G. GEVERT, S. WANG and R. U. BROZE: Heat to a.c. conversion using thermoelectric generators and tunnel diodes

I. BRAUN and E. Y. WANG: A new method of zone leveling for materials having distribution coefficients in the range of 0–0·6

J. R. DALE and J. C. BRICE: Etch pits in germanium and their relation to hardness

G. STRULL and W. M. KAUFMAN: An astable four-region multivibrator

J. M. GOLDEY, I. M. MACKINTOSH and I. M. ROSS: Turn-off gain in p–n–p–n triodes

T. E. HARTMAN: Transient photovoltaic response of diffused-junction silicon photodiodes

Vol. 2
1961